Extreme metabolomics

Jankevics, Andris

Extreme Metabolomics

Developing a high-performance computational pipeline for high-resolution LC–MS data sets

The work described in this thesis was carried out at the Groningen Bioinformatics Centre, University of Groningen, The Netherlands and at the Institute of Molecular, Cell and Systems Biology, College of Medical, Veterinary and Life Sciences, University of Glasgow, United Kingdom. The research was financially supported by an NWO-Vidi award to Prof. Dr. Rainer Breitling.

The author gratefully acknowledges the financial support of the University of Groningen and the Groningen Biomolecular Sciences and Biotechnology Institute (GBB) for printing this thesis.

Artwork	Cyclist illustration by Francesco Rollandin, obtained from `www.openclipart.org`. Fragment of KEGG global pathway map used on back cover and bookmark was generated using IPath tools (Kanehisa *et al.*, 2012; Yamada *et al.*, 2011).
Printed by	Minuteman Press, Stockport, United Kingdom
ISBN	978-90-367-6310-3 (printed version)
ISBN	978-90-367-6309-7 (digital version)

RIJKSUNIVERSITEIT GRONINGEN

Extreme Metabolomics

Developing a high-performance computational pipeline for high-resolution LC–MS data sets

Proefschrift

ter verkrijging van het doctoraat in de
Wiskunde en Natuurwetenschappen
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, dr. E. Sterken,
in het openbaar te verdedigen op
vrijdag 27 september 2013
om 12.45 uur

door

Andris Jankevics

geboren op 11 mei 1981
te Riga, Letland

Promotores :	Prof. dr. R. Breitling
	Prof. dr. R.C. Jansen
Beoordelingscommissie :	Prof. dr. M. Heinemann
	Prof. dr. R. Goodacre
	Prof. dr. J.A. Westerhuis

Contents

Abstract

Metabolomics aims at understanding biology by comprehensive metabolite profiling (the simultaneous measurement of as many low molecular weight compounds as possible in a biological sample). A favourite method for this kind of profiling is Liquid Chromatography coupled to Mass Spectrometry (LC–MS).
The goal of the work described in this thesis was to develop open source software for the analysis of LC–MS metabolomics data. A highly modular design of the software allowed us not only to fine-tune the analytical components relevant for each specific study design, but also to share intermediate analysis results with the biologists and analytical chemists on the project in an intuitive way. This new approach facilitated the communication between researchers with different backgrounds and resulted in the rapid discovery of bottlenecks in experimental design, bioanalytical methodology and data analysis.
The thesis starts by introducing the benefits of metabolomics using a mass spectrometry platform and an overview of the entire experimental pipeline (**Chapter 1**). In **Chapters 2–6** various advanced data filtering and metabolite identification techniques are described. In **Chapter 7** we are giving a working example of the flexible computational workflow for data processing.
These methodological chapters are followed by several biological studies applying metabolomics in a systems biology context. The optimisation of a comprehensive metabolite extraction protocol for *Leishmania donovani* parasites and the subsequent optimisation of the analytical approach is described in **Chapter 8**. In **Chapter 9** we are able to show that the overexpression of an antisense non-coding RNA targeting glutamine synthetase I results in a major reorganization of the metabolism of the bacterium *Streptomyces coelicolor*, by applying metabolomics and a new computational approach based on concordance analysis to an extremely large number of analytical replicates. **Chapter 10** demonstrates the application of stable isotope labelling and untargeted metabolomics to obtain a global overview of the cellular fate of precursor metabolites.
Finally, the thesis concludes in **Chapter 11** where we outline future perspectives for further research in metabolomics.

Chapter 1

Introduction

This chapter is a modified version of:

Maya Berg[1], Manu Vanaerschot[1], Andris Jankevics[2,3,4], Bart Cuypers[1], Rainer Breitling[2,3,4], Jean-Claude Dujardin[1,5]

Comput. Struct. Biotech. J. 4(5):e201301002, 2013

1 Unit of Molecular Parasitology, Department of Biomedical Sciences, Institute of Tropical Medicine, Antwerp, Belgium
2 Institute of Molecular, Cell and Systems Biology, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, United Kingdom
3 Groningen Bioinformatics Centre, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Groningen, The Netherlands
4 Faculty of Life Sciences, Manchester Institute of Biotechnology, University of Manchester, Manchester, United Kingdom
5 Department of Biomedical Sciences, University of Antwerp, Antwerp, Belgium

The favorite technology for global metabolic profiling (metabolomics) are so-called hyphenated mass spectrometry (MS) platforms, such as gas chromatography–mass spectrometry (GC–MS), liquid chromatography–mass spectrometry (LC–MS) or capillary electrophoresis–mass spectrometry (CE-MS) (Halket *et al.*, 2005). Alternatively, NMR spectroscopy, direct infusion atmospheric pressure ionization (API) MS, and other methods, such as Raman spectroscopy and Fourier transform infrared spectroscopy, can be used for higher throughput but less specific metabolomics screening experiments (fingerprinting) (for a comparison see (Scheltema *et al.*, 2010)). The selection of the platform is always a compromise between sensitivity, speed and chemical selectivity and coverage of the relevant subset of the metabolome (Boccard *et al.*, 2010). One must bear in mind that the chemical diversity and the range of concentration of different metabolites is very diverse, therefore no single platform provides a complete coverage of the metabolome (Creek *et al.*, 2012b).

Chromatographic separation by LC or GC has two main advantages when compared to direct-infusion MS: (i) it separates isomers (metabolites of a same mass) which would appear as indistinguishable entities in downstream MS analysis; and (ii) it minimizes ion suppression in which a more easily ionizable species masks the presence of a less ionizable one (Lakshmanan *et al.*, 2011) hence allowing a higher quantitative accuracy (Zheng *et al.*, 2010). The hyphenation of MS, i.e. its combination with a chromatographic separation, greatly increases the quality of the raw data generated and the number of metabolites to be detected, but it also increases the analysis time (Boccard *et al.*, 2010). A detailed comparison between GC–MS and LC–MS — the two main separation methods in metabolomics — is described elsewhere (Creek *et al.*, 2012b; Halket *et al.*, 2005). In short, compared to LC–MS, GC–MS analysis involves a more complex sample preparation, since it is only capable of analyzing volatile compounds or those that can be made volatile by derivatization (Dunn *et al.*, 2011a). In addition, many polar compounds are not detectable by GC–MS, and due to the electron ionization (EI) technique used in GC–MS, only the most abundant positively charged ions are measured (Creek *et al.*, 2012b; Halket *et al.*, 2005).

However, GC–MS generates reproducible fragmentation patterns, for which fragment databases exist (that can be shared between investigators), and produces stable retention times, which can be matched with existing libraries containing retention time information, for a huge array of analytes Scheltema *et al.* (2010). This makes it much easier to verify the identification of detected metabolites in GC–MS. The LC–MS situation is more complex: the atmospheric pressure ionization techniques (APCI, ESI) produce both positively and negatively charged ions, but suffer more

from matrix effects and ionization suppression or enhancement. In addition, while LC–MS also generates characteristic retention times for each metabolite, which assists in metabolite identification, these retention times are more difficult to reproduce and compare between laboratories and library matching is still at an early stage. In reality, choosing between a GC–MS and an LC–MS platform is most often determined by the availability of a platform and existing collaborations.

1.1 Data analysis

Despite recent advances in methodology, metabolomics still presents a number of challenges, including both technological issues and limitations of data interpretation (Kind and Fiehn, 2009). As typically a large amount of signals is detected, data complexity usually is so high that it is not possible to interpret data manually; hence, specific software tools and algorithms are needed. These types of analysis also require fast processors and huge storage capacity, typically in the terabyte range for large datasets. Comprehensive overviews of many of the existing tools for data processing in metabolomics have been presented recently (Katajamaa and Orešič, 2005; Melamud *et al.*, 2010; Prince and Marcotte, 2006; Scalbert *et al.*, 2009; Zelena *et al.*, 2009).

The first part of this thesis (**Chapter 2–7**) is describing our developments and improvements in data processing and data handling routines. General data processing steps include feature or peak detection (Tautenhahn *et al.*, 2008), peak matching and several additional steps of signal filtering and noise removal (Windig, 2004). For example, the peak matching step involves aligning of the chromatographic features between technical or biological replicates of a single sample. Peaks that are not detected in all technical replicates can be discarded from further analysis. Derivative signals such as isotopes, adducts, dimers and fragments, can be automatically annotated by correlation analysis on both signal shape and intensity patterns using software tools like CAMERA (Kuhl *et al.*, 2012), PUTMEDID-LCMS (Brown *et al.*, 2011) and mzMatch (presented in detail in **Chapter 2**). Such peaks are not discarded, but only flagged, so that their assigned annotations can be taken into account in the metabolite identification step.

Metabolite identification in LC–MS is mainly based on matching the detected mass with available mass databases and derivative signal annotations. In contrast to proteomics, efficient algorithms that can reasonably successfully predict and compare the mass fragmentation patterns for tandem MS spectra of metabolites (e.g. MetFrag (Wolf *et al.*, 2010)) still need further development. In addition, in **Chapter 3**

we suggest applying a quantitative structure retention relationship (QSRR) model on authentic standard compounds to predict retention times of chemically similar compounds (Creek *et al.*, 2011). Including the predicted retention times in the identification step significantly improved metabolite identification by removing 40% of falsely identified compounds, which had the correct mass but inconsistent retention time.

As the nature of metabolic experiments and experimental design varies widely, there is a demand for software tools that could easily be adapted for the evolving demands of data processing, for example, adding extra data filtering tools or changing the order of a typical processing pipeline. **Chapter 4, 5** and **7–10** illustrates the configurability of the mzMatch software toolkit and it's extension for the complete processing of raw mass spectra, including steps for noise filtering and compound identification by matching mass databases. The PeakML file format used by mzMatch allows users to share data with other commonly used software packages such as XCMS (Smith *et al.*, 2006), mzMine (Pluskal *et al.*, 2010b) and IDEOM (Creek *et al.*, 2012a), giving flexible access to an extended set of data processing tools. For instance, IDEOM (**Chapter 6**) is a user-friendly Excel interface to mzMatch, which allows researchers to run a comprehensive pipeline for data- analysis and visualization from a graphical user interface within Microsoft Excel. Efforts to develop such tool chains and (semi-) automated tools for data processing in unified data exploration platform were a primary goal of this thesis and will be a priority in the near future.

1.2 Current issues with LC–MS

A major drawback of LC–MS is that it only allows for semi-quantitative analysis. For example, LC–MS signals do not always scale linearly with metabolite concentrations, as has clearly been shown by dilution series in **Chapter 4**. The deviation from linearity strongly depends on (i) the type of column (HILIC versus C18; HILIC is more prone to variations in signal intensity); (ii) the concentration of the metabolite (ion suppression occurs more frequently with higher concentrations); and (iii) the loading capacities of the column. Therefore, it is important to always interpret the metabolite profiles in terms of relative quantification, where the raw peak height of a metabolite of interest is compared to the raw peak height of the *same* metabolite in a reference sample or, for example, other samples in a time series. During the last few years, however, absolute quantification of selected compounds using ^{13}C-labelled standards is gaining ground in global metabolomics studies (Creek *et al.*, 2012b,c; Scheltema *et al.*, 2010), providing unique insights into the dynamics

of metabolic fluxes, beyond the steady-state information gathered by routine mass spectrometry (Creek *et al.*, 2012c). The Maven (Melamud *et al.*, 2010) and mzMatch-ISO (**Chapter 5**) packages can be used to process isotope-labeled data sets [`http://mzmatch.sourceforge.net/untargeted_labelling.php`]. In **Chapter 10** we give an example of the application of stable isotope labelling and untargeted metabolomics to obtain a global overview of the cellular fate of precursor metabolites.

The combination of LC–MS based metabolomics data that were collected over a longer period of time on the same platform and in the same laboratory remains problematic due to the systematic variability between LC–MS measurements (Zelena *et al.*, 2009). This systematic variability includes variable ionization (influenced by many factors, such as co-elution of other metabolites, salts, pH of the mobile phase etc.), drift in retention time (column degradation or replacement), and drift in mass calibration (changes in temperature and electronic circuitry). Samples that need to be compared should therefore preferably be measured in the same run (and if possible in randomized order), although total run length should be limited, because contamination will cause drifts in the measured response and retention time over relatively short analysis periods (tens of injections). These drifts in both retention time and mass accuracy are detrimental for platforms such as LC–MS that depend on these parameters for identification (Dunn *et al.*, 2011a). A drift in retention time can occur when a large number of metabolite extracts is measured in one run on the LC–MS platform or during different analytical blocks that need to be pooled afterwards. This can result in sets of peaks of a single metabolite being considered as belonging to two different compounds in the peak matching step. Re-alignment of the retention time over different samples with the OBI-Warp tool (Prince and Marcotte, 2006) followed by gap-filling (secondary peak picking step to retrieve missing signals within a specified retention time and mass window from the raw data files) (Katajamaa and Orešič, 2005) can be applied and will significantly reduce the number of double identifications in the eventual list of identified compounds. To handle drift in mass calibration, ubiquitously detected contaminants of known exact mass can be used for internal mass calibration or to align spectra after the unavoidable mass drift during long term studies (Breitling *et al.*, 2012). One can also include replicate measurements of a series of authentic standards (e.g. the ones used for the above mentioned QSSR model) covering the whole mass range of interest, which will allow recalibrating during data processing.

When large metabolomics studies divided over a series of analytical blocks cannot be avoided, normalization of the data can be considered. Dunn *et al.* (2011a) suggest using a standard quality control sample representative of the sample type under

analysis to allow for signal correction within and between analytical blocks. This kind of normalization is model-driven, where an external model is extrapolated to the dataset of interest.

1.3 Designing metabolomics experiments

Many components of the metabolome change only very slowly throughout life, making selected metabolites popular biomarkers in human medicine (cholesterol being a very common example). However, at the cellular level changes can be much more rapid, and between different cell types or developmental stages, large fractions of the metabolome can be drastically rearranged. This has important consequences for the choice of sample for a metabolomics analysis. Many unicellular pathogens, especially those that are transmitted through a vector, have different life forms. Choosing the correct life stage of the organism under study greatly depends on the research question. Are the suspected differences expected to be present throughout all life stages or only at one specific stage? In the case of *Leishmania* for example, the intracellular amastigotes are the most clinically relevant form to study, as only this form occurs in the human host. However, amastigotes have as yet not been thoroughly studied at the metabolomics level due to several technical constraints (difficulty to separate its metabolome from that of the host cell, quick transformation to promastigote life stage upon isolation, difficulty of obtaining sufficient quantities) (Decuypere *et al.*, 2008). Free-living pathogens, such as trypanosomes belonging to the subgenus *Salivaria*, create fewer problems concerning the choice of life stage for metabolomics studies, since both the procyclic (fly vector) and the bloodstream form (human host) can be easily extracted (Kamleh *et al.*, 2008; Vincent *et al.*, 2012). The extracellular promastigote form of *Leishmania*, which naturally occurs in the vector, is easier to culture *in vitro* and is therefore also the most studied life form of the parasite in metabolomics and other studies.

Another issue often affecting metabolomics studies is that cells come in different sizes: when comparing the metabolic profile of two samples with a significant difference in cell size, the eventual results can be skewed, with the larger cell showing generally increased metabolite levels which is superimposed upon the metabolite changes of interest. In contrast to, for example, transcriptomics, no commonly accepted standard procedure is available for correcting this bias. Normalizing the metabolomics results according to the cell size might be recommendable if such differences are known to occur. Although such a normalization method may seem justified to biologists, many LC–MS specialists feel that this is perilous because LC–

MS signals do not always scale linearly. The semi-quantitative nature of LC–MS measurements allows only comparison of the same (!) metabolites between different samples within the same measurement block, and not comparison of the quantity of different metabolites within a given sample. Hence, one single normalization factor for all metabolites could over- or under-correct the intensities of metabolites with different physicochemical properties. Nevertheless, protein content normalization has already been applied when comparing the metabolic profile of one *Leishmania* strain at different stages in the promastigote growth curve (Silva *et al.*, 2011). In this study, it was shown that transformation to the metacyclic form (the smaller infective form) was accompanied by a decrease in protein content, which is thought to correlate with the decrease in cell size. Hence, by determination of the total protein content present in a sample with a commercially available kit, differences in cell size can be corrected.

1.3.1 Sampling

A sampling protocol that minimizes the biological and technical variability is indispensable for any biological metabolite profiling study. This is also the case for metabolomics, since the metabolome can change very rapidly, for example in response to differentiation processes or subtle changes in the environment (such as temperature fluctuation, osmotic stress, or nutrient depletion). Thorough preparation of the whole sampling pipeline will be imperative to ensure a swift sample preparation that minimizes the induction of additional biological or technical variability induced by the sampling procedure itself. To reduce the technical variability throughout the sampling procedure, it is of utmost importance to bring the metabolism of the cells rapidly to a halt and avoid leaking of metabolites during the various washing steps before the actual metabolite extraction. An example of sampling protocol optimisation of a unicellular pathogen, *Leishmania donovani*, is given in **Chapter 8**. In **Chapter 9** we demonstrate the power of highly replicated experimental designs for the robust characterization of metabolite dynamics in a Gram-positive bacterium, *Streptomyces coelicolor*, despite the well-known biological variability between batch cultures.

In addition to various control samples, and depending on the study outline, other samples can be obtained simultaneously as well. If a parallel genomic or proteomic study is planned, for example, it is best to prepare these samples at the same time as the metabolomics study, due to the variability of the metabolome and the plasticity of the *Leishmania* genome (Downing *et al.*, 2011; Mannaert *et al.*, 2012). This significantly facilitates later integration of the metabolome and the genome, transcriptome and/or proteome results into a general systems biology interpretation. Furthermore,

when processing several strains together, the genome sequence can also be used as a quality control to confirm the identity of the material used.

1.3.2 Sample storage

Just before storage of the metabolomics samples at –80 °C, samples should be deoxygenated with a gentle stream of nitrogen gas for 1 min prior to tube/vial closure (t'Kindt *et al.*, 2010b). The effect of storing serum and urine samples at 4 °C for 0 h or 24 h prior to storage at –80 °C has been shown to be small: the observed variance between samples due to storage at 4 °C for 24 h was of the same magnitude as the analytical variance associated with replicate analysis per sample (Dunn *et al.*, 2008). Dunn *et al.* (2011a) recommend analyzing samples within 2 years of sample collection and avoiding multiple freeze-thawing cycles of a single aliquot. Optimally, a sample should be opened only once and, if needed, multiple aliquots of the same subject can be collected (Dunn *et al.*, 2011a). To our knowledge, reports on the stability of metabolites present in parasite samples, or plasma, serum, urine or cerebrospinal fluid samples for that matter are rather scarce (Dunn *et al.*, 2008, 2011a; Gika *et al.*, 2007; Rosenling *et al.*, 2011).

1.3.3 Sample list set up

The decision which samples to measure by LC–MS and in which order, is far from trivial, as it affects the accurate assessment of biological and technical variability. Most metabolomics studies include three or four biological replicates of each experimental treatment (Creek *et al.*, 2011; De Souza *et al.*, 2006; Jankevics *et al.*, 2011; Silva *et al.*, 2011; t'Kindt *et al.*, 2010a). Beside the biological replicates, a series of other samples should be included in each LC–MS run. First of all, a reference sample should be injected at least four to eight times to equilibrate the analytical platform and assess the reproducibility of subsequent runs (Gika *et al.*, 2008). Preferably, the reference sample is similar to the actual samples of interest in complexity and composition. In addition, commercially available authentic standards can be added to the measurement series to compare retention time for metabolite identification. For example, in **Chapter 3** we describe the use of 2 mixtures of authentic standards (127 metabolites in total), which can be used to predict retention times also for compounds that are not included in the mixture (Creek *et al.*, 2011). Finally, a dilution series of a pooled sample of all extracts can be included (**Chapter 4**), which will help to filter out a substantial part of spurious signals (Jankevics *et al.*, 2012). The measurements of both the standards and the reference samples should be regu-

larly distributed throughout the sample list so they can be used for quality control to assess LC–MS stability (Dunn *et al.*, 2011a; Sangster *et al.*, 2006). Additional controls can include cell-free growth medium and extraction solvent blanks to filter out contaminant peaks by "blank" subtraction. The order of all these samples should be well considered: randomization of the different samples within blocks of four biological replicates alternated with quality control samples is recommended; this will allow detecting systematic variability throughout the LC–MS measurement (t'Kindt *et al.*, 2010a). Figure 1.1 illustrates a recommended sample sequence, based on a dilution series of quality control samples and randomization of analytical samples within blocks (shown for two biological replicates per sample). For example, if all biological replicates of condition 1 are measured first, followed by all biological replicates of condition 2, technical issues during the LC–MS experiment (in particular the unavoidable column degradation) would result in a confounding of experimental and temporal factors, seriously interfering with the later statistical interpretation of the data. By randomizing the biological replicates in a well-considered way, this can be largely avoided (Figure 1.1). The quality control samples which alternate with the biological replicates are used to detect and potentially correct for these technical issues.

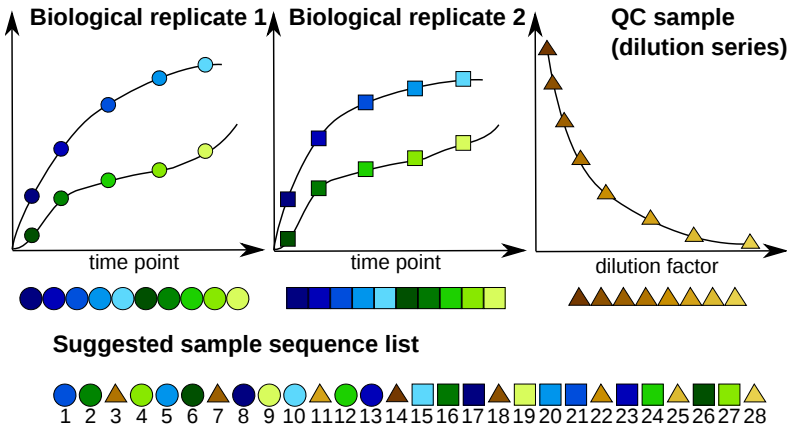


Figure 1.1: Recommended sample sequence for samples from two different experimental conditions (condition 1 = shades of blue, condition 2 = shades of green) measured at five different time points. Each condition has two biological replicates. The y-axis represents the measured intensities of the biological replicate, whereas the x-axis represents the five different time points. Samples from a dilution series of a quality control pooled reference samples (shades of brown) are interspersed at regular intervals in the suggested sample sequence.

1.3.4 LC–MS measurement

The most widely used liquid chromatography system in metabolomic research on *Leishmania* and related protozoan parasites (Kamleh *et al.*, 2008; Silva *et al.*, 2011; t'Kindt *et al.*, 2010a,b) is the HILIC column (hydrophilic interaction liquid chromatography): it allows polar metabolites to be retained, whereas lipophilic metabolites elute relatively rapidly from the column. This is a significant advantage compared to reversed-phase columns, from which lipids are difficult to elute and can accumulate and cause ion suppression by their background bleed (Kamleh *et al.*, 2008). The performance of two HILIC columns with a different inner diameter (2.1 mm versus 4.6 mm) has been compared (**Chapter 8**) and showed that the number of putatively identified metabolites dropped nearly two-fold for the wider column (from 390 to 220). Using the 2.1 mm HILIC column, 20% of the predicted metabolome of *Leishmania* could be detected (t'Kindt *et al.*, 2010b). However, it was also notable that the narrower column behaved in a less reproducible way, especially in terms of retention time drift.

For large batch analysis, the 4.6 mm HILIC might thus be the preferable column. To obtain a more complete coverage of the metabolome, measurements on different types of columns can be combined. For example, since lipids are considered to be biologically important in *Leishmania* drug metabolism (t'Kindt *et al.*, 2010a), lipidomics studies in *Leishmania* are becoming of increasing importance to unravel the mechanisms of drug resistance (Imbert *et al.*, 2012). These studies use different types of columns, ranging from HILIC (Zheng *et al.*, 2010) to normal phase (Imbert *et al.*, 2012).

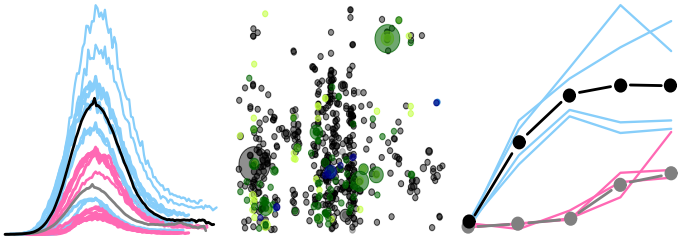
For an overview of the existing ion separation methods we refer to Watson (2010). In summary, the Orbitrap mass spectrometer is the most sensitive instrument currently applied in general metabolomics studies: it combines ultra-high mass accuracy (<1 ppm) and resolution (>100,000) with a high dynamic range (approx. 10^5), allowing unambiguous assignment of a molecular formula to many observed masses (Creek *et al.*, 2012b). LC–MS generally analyzes samples in both positive (ESI+) and negative ion mode (ESI–), as they provide complementary data. The Orbitrap Exactive configuration is well-suited for this kind of analysis, since it has a positive–negative polarity switch mode which reduces analysis time, amount of sample needed and issues related to combining the two modes afterwards if they were not recorded simultaneously (such as retention time drif). Time-of-Flight instruments (TOF) are also compatible with chromatographic systems interfaced to an ESI source, but linear dynamic ranges are around 10^3 and the resolving power is limited to 3 ppm (Watson, 2010).

Chapter 2

PeakML/mzMatch: a file format, Java library, R library, and tool-chain for mass spectrometry data analysis

Richard A. Scheltema[1*], Andris Jankevics[1,2*], Ritsert C. Jansen[1], Morris A. Swertz[1,3], Rainer Breitling[1,2]

Anal. Chem. 83(7):2786–93, 2011.



1 Groningen Bioinformatics Center, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Groningen, The Netherlands
2 Institute of Molecular, Cell & Systems Biology, College of Medical, Veterinary & Life Sciences, University of Glasgow, Glasgow, United Kingdom
3 Genomics Coordination Center, Department of Genetics, University Medical Center Groningen and University of Groningen, Groningen, The Netherlands.
* Equal contribution

abstract

The recent proliferation of high-resolution mass spectrometers has generated a wealth of new data analysis methods. However, flexible integration of these methods into configurations best suited to the research question is hampered by heterogeneous file formats and monolithic software development. The mzXML, mzData, and mzML file formats have enabled uniform access to unprocessed raw data. In this paper we present our efforts to produce an equally simple and powerful format, PeakML, to uniformly exchange processed intermediary and result data. To demonstrate the versatility of PeakML, we have developed an open source Java toolkit for processing, filtering, and annotating mass spectra in a customizable pipeline (mzMatch), as well as a user-friendly data visualization environment (PeakML Viewer). The PeakML format in particular enables the flexible exchange of processed data between software created by different groups or companies, as we illustrate by providing a PeakML-based integration of the widely used XCMS package with mzMatch data processing tools. As an added advantage, downstream analysis can benefit from direct access to the full mass trace information underlying summarized mass spectrometry results, providing the user with the means to rapidly verify results. The PeakML/mzMatch software is freely available at http://mzmatch.sourceforge.net, with documentation, tutorials, and a community forum.

Recent years have seen new and exciting metabolomics and proteomics experiments enabled by an increasing variety and improved performance of mass spectrometry equipment (Dunn, 2008; Han *et al.*, 2008). Standardization initiatives such as the mzXML file format (Pedrioli *et al.*, 2004) and the recently introduced mzML (Deutsch, 2010) have considerably shortened the development cycle for analysis software, and, as a consequence, a wealth of data analysis applications has become available, of which XCMS (Smith *et al.*, 2006) and mzMine (Pluskal *et al.*, 2010b) are commonly used for metabolomics data sets. However, mzXML only standardizes the description of raw mass spectrometry data. Downstream integration of analysis tools into suitable analysis pipelines is still hindered by (i) use of monolithic, black box approaches where algorithms and resulting (intermediate) data are inaccessible to the user for verification, (ii) limited ability to check intermediate results of data processing and limited access to the underlying context information e.g., peak shape or neighboring peaks), and (iii) data format heterogeneity at various steps of the analytic pipeline, making it difficult to recombine pipeline components to suit new experiments or technologies, e.g., when changes in chromatographic conditions or

mass accuracy require different peak picking or filtering modules.

In the literature, a number of initiatives to standardize the storage of extracted features have already been described, including the initiatives developing FeatureXML (Sturm and Kohlbacher, 2009) and CMLspect (Kuhn *et al.*, 2007). Even though these provide basic support for storage of extracted chromatogram information, they lack many of the options that PeakML offers. Additionally, some of their features, such as FeatureXML's comprehensive framework for storing protein/peptide identifications, create unnecessary overhead and restrictions in the context of metabolomics experiments.

In order to pick up where the current open formats leave off, we have developed the PeakML file format, an open and extensible format for the standardized representation of peak and meta-information from each step in a downstream analysis pipeline. The power of PeakML and mzMatch for rapid tool integration is demonstrated by a collection of small tools (Scheltema *et al.*, 2008, 2009) and the availability of PeakML read and write functionality for XCMS (Smith *et al.*, 2006), a widely used data analysis software (Arbona *et al.*, 2009; Dai *et al.*, 2010; Lin *et al.*, 2009). Equivalent converters can easily be created for other generic mass spectrometry processing tools. This comprehensive collection of components is intended to further encourage a modular and interchangeable design of analysis components, storing data generated/extracted by each step in a standardized manner. The added value for algorithm developers is that they can build on off-the-shelf PeakML software components (e.g., for data loading and visualization) and gain access to a potentially much larger user community for their tools.

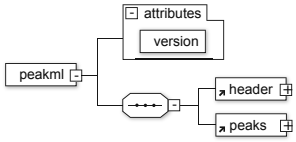


Figure 2.1: PeakML file format. The format consists of two separate blocks: the *header* and the *peaks* block.

2.1 Application programming interface

2.1.1 PeakML file format

The PeakML file format was specifically designed to provide an open XML standard for the storage of hyphenated mass spectrometry data, of which a summary is given here. It differs from existing efforts such as mzML in that it supports common downstream results, ranging from storing mass chromatograms, background ions (Scheltema *et al.*, 2008), and any combination of either. In order to support the development of automatic processing software, a comprehensive metadata structure is provided. Information about the measurements (such as experimental parameters and machine settings) and how they are organized can be stored in this structure. The XML schema of the PeakML file format can be divided into two blocks (Figure 2.1): a *header* block for storing the metadata and a *peaks* block for storing peak information.

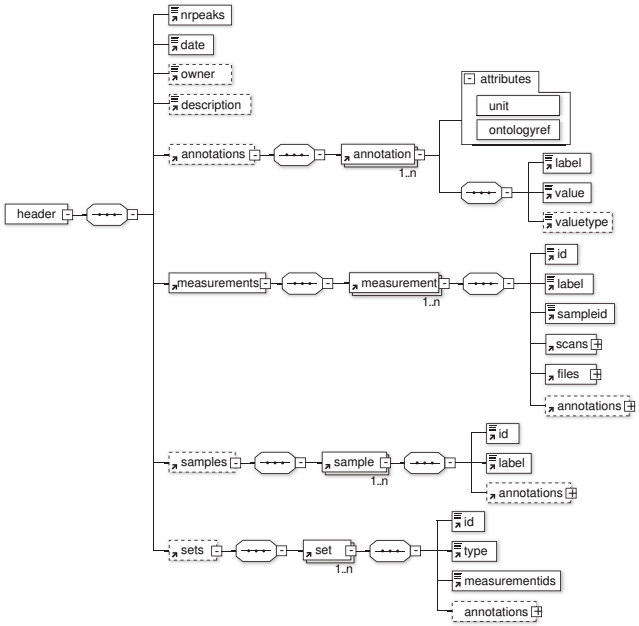


Figure 2.2: The *header* block stores general information about the contents of the file (e.g., date of creation). Beside this general information, full descriptions on the measurements are provided (e.g., such as ionization mode, original file, etc.) and how they are organized in sets. Each measurement also contains a measurement-id, which can be used to link the measurement data to the information stored in the *peaks* block. The contents of each section can be extended with annotations (label-value pairs).

The *header* block (Figure 2.2) provides the structure for storing the required metadata for each step divided into four components (measurements, samples, sets, and applications), each of which can be extended with annotations supporting controlled vocabularies (Bodenreider, 2004; Stoeckert and Parkinson, 2003; Whetzel

et al., 2006). (i) The obligatory *measurements* block contains basic information about the measurements collected in the current analysis, such as names of all the files used, and a link to the sample description. (ii) The *samples* block contains information on the sample, which is kept to a basic level with an id, optional label, and annotations. (iii) Multiple measurements can be organized in a set or set of sets, e.g., all technical replicates for a certain sample are part of a single set, and the technical replicates for several related biological replicates are organized as a set of sets. The *sets* block contains information on how the measurements are organized in these sets, which can be used by automatic processing software, such as an RSD filter (Shah *et al.*, 2000) that filters nonreproducible peaks across technical and/or biological replicates, and by visualization software, that assigns a single color to all peaks in a set of technical or biological replicates. (iv) The *applications* block contains information about the software components or "steps" used to produce the file, e.g., a peak extraction tool would provide information such as its version, the raw data file, and the expected mass accuracy of the machine in parts-per-million. This allows for a complete reconstruction of the entire analysis protocol used and provides an archival trace of all raw and intermediate data files used to generate the current data set.

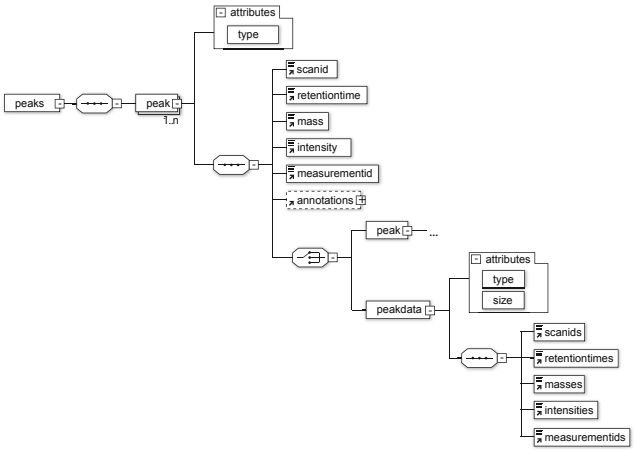


Figure 2.3: The *peaks* block holds all intermediate and result mass spectrometry data. A peak is described by its mass, intensity, measurement-id, optional scan (for LC–MS data), and optional annotations for a peak. Each peak can be typed as being either "backgroundion" or "mass-chromatogram" (using the peakdata element), or as type "peakset" (using the recursive *peak* element).

The *peaks* block (Figure 2.3) provides the complete structure for storing information on one or multiple peaks. As the file format is focused on hyphenated mass spectrometry data, a peak is defined here as either a mass chromatogram, a background ion, or a set of one of these. Background ions (Scheltema *et al.*, 2008) in PeakML are defined as analytes present over the whole retention time range (and are generally of no interest for the biological interpretation), while mass chromatograms (EICs) are caused by analytes eluting over a narrow retention time window (and are

of interest for the biological interpretation).

The type of each peak entry is identified by the attribute type (which supports: *masschromatogram*, *backgroundion*, and *peakset*), providing an extensible construct for future versions and backward compatibility. Each entry is opened with summary information: "scanid", "retentiontime", "mass", "intensity", and "measurementid", which can be used to load the entries as a flat peak table, without the complete trace information. More annotations can be stored for each peak entry, analogous to the header entries. The real data for each peak entry is made up of either a new *peak* entry (only for type "peakset") or a *peakdata* entry containing the trace information (for "masschromatogram" and "backgroundion"). The trace information is stored as Base-64-encoded arrays in little-endian ordering (analogous to the mzXML and mzML formats), reducing the memory requirements. Currently only centroid single mass analyzer mode data are supported, but facilities for profile data are in place for future support. Support for additional mass spectrometry data, including tandem spectra, lies outside the aim of this format, as this is covered by other data formats (Eisenacher, 2011).

In order to minimize nongeneralizable, application-specific overhead (e.g., protein identification support) in the format, we decided to focus solely on extracted feature support. The concept of annotations provides an extensible platform to add additional identification information, for which two approaches can be taken: (1) Storage of the most pertinent information itself, e.g., as a descriptive text, or (2) storage of a link to an entry in a companion file containing the complete information (e.g., protein identification stored in an mzIdentML (Eisenacher, 2011) file).

PeakML allows the effective storage of mass chromatograms (extracted ion chromatograms) for single or multiple measurements, achieving an average data reduction of around 75%. Typical file sizes for input data in mzXML format and PeakML data files are shown in Table 2.1. Examples of PeakML files are available for download at `http://mzmatch.sourceforge.net/peakml_files.html`.

We envision that PeakML will be used in combination with other formats that describe the whole experimental process (protocols, biomaterials, experimental variables, hypotheses, and conclusions). These are the domain of complementary data models such as FuGE (Jones and Lister, 2010), MIAPE (Taylor *et al.*, 2007) and XGAP (Swertz *et al.*, 2010).

Sample[a]		mzXML data (Mb)	PeakML data (Mb)			Number of features or peak sets in PeakML file		
Cond.	**Rep.**		**1**	**2**	**3**	**1**	**2**	**3**
1	A	10.4	2.0			9504		
1	B	12.3	2.3	9.5	25.8	10696	16909	23926
1	C	12.0	2.5			11867		
2	A	10.1	2.1			10027		
2	B	12.0	2.3	9.7	25.8	10897	17556	23926
2	C	12.2	2.5			12014		
3	A	10.9	2.3			10865		
3	B	12.1	2.1	8.7	25.8	9996	15131	23926
3	C	10.1	1.8			8742		

Table 2.1: File size comparison. a) Samples were acquired in centroid mode on a Thermo LTQ-Orbitrap XL instrument. Binary native data were converted to the mzXML file format with ReAdW (a tool of the Trans-Proteomic Pipeline software collection, downloaded from `http://tools.proteomecenter.org/wiki/index.php?title=Software:ReAdW`). A final reduction of file size by 75% is achieved. Cond. – analytical condition; Rep. – biological replicate, 1 – PeakML file of the single LC–MS run after peak detection, 2 – PeakML file of the combined peak sets of biological replicates, 3 – PeakML file of combined peak sets between biological replicates and conditions.

2.1.2 The PeakML library

The PeakML Java library defines the fully documented Application Programming Interface (API) for handling PeakML data, parsers to load mass spectrometry files, and other components useful for building mass spectrometry analysis tools, including (but not limited to) chemistry, math, and user interface routines, considerably speeding up application development.

The core of the library consists of classes specific to easily interact with mass spectrometry data stored in raw and PeakML file formats (see Figure 2.4). The base class *IPeak* defines the minimal properties of each data element, such as mass, intensity, scan-number, and retention time; additionally a measurement-id can be defined to link an instance of *IPeak* to a *Measurement*. From *IPeak* commonly encountered mass spectrometry data types are extended including *Spectrum*, *MassChromatogram*, *BackgroundIon*, *ChromatographyMS*, and *PeakSet*. The class *Spectrum* defines a single scan of a mass spectrometer, either in continuous or centroid mode. The classes *MassChromatogram* and *BackgroundIon* represent a mass trace of an analyte in a hyphenated setup (i.e., GC–MS or LC–MS), extracted from a set of consecutive spectra. All *MassChromogram* objects from one experiment can be stored as a bundle in the class *ChromatographyMS*. Mass spectrometry data sets are

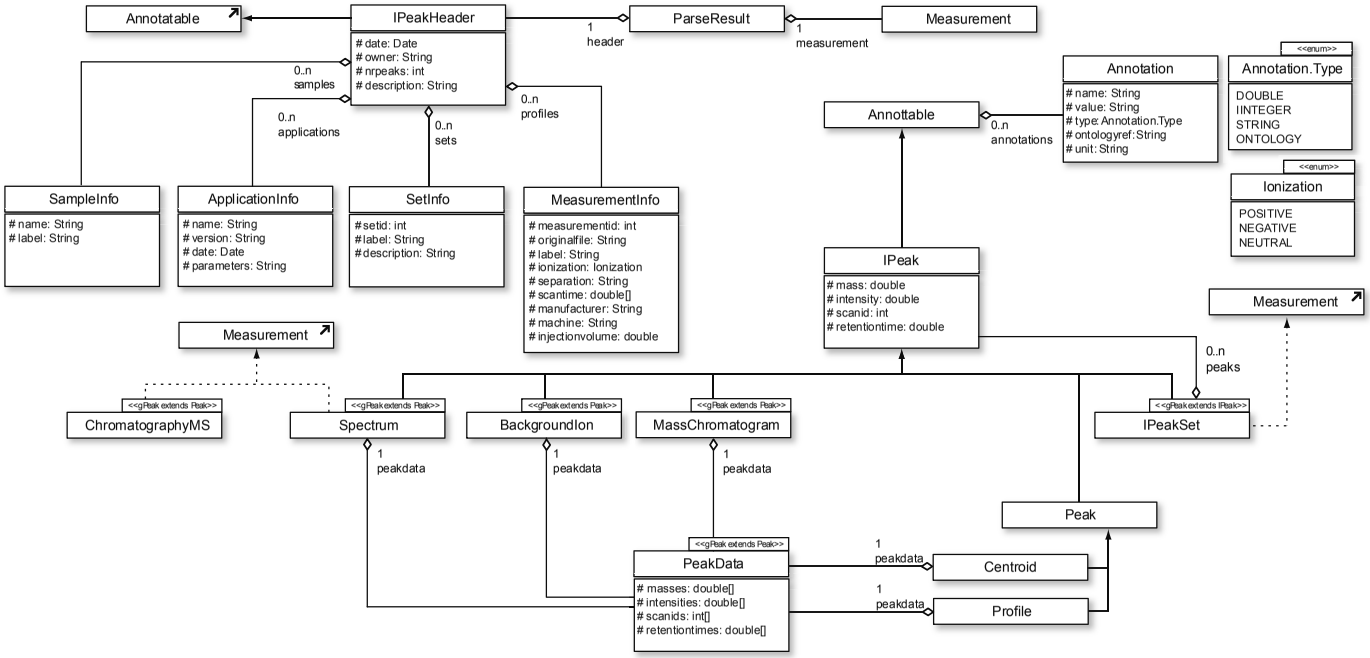


Figure 2.4: UML diagram of the mass spectrometry specific interface. The diagram contains two subsets, the mass spectrometry classes and the I/O classes. (1) The base-class is *IPeak*, which defines the basic properties of a signal or peak (mass, intensity, scan-id, and retention time, plus a measurement-id to link it to meta-information on the measurement). These properties are further specialized and extended in the subclasses (e.g., retention time for a mass chromatogram is defined as the time where the signal is most intense). The subclass *IPeakSet* copies and extends on the behavior of *java.util.Vector* by providing additional functionality on the list of peaks (e.g., fast binary search on *IPeak* properties). (2) Each parse method returns an instance of *ParseResult*, containing pointers to a *Header* instance and an instance of type *Measurement* (either *Spectrum*, *ChromatographyMS*, or *IPeakSet*), which was introduced to identify those classes that can be stored in a file. The class *Annotatable* (arguably the real base-class) injects functionality into classes for adding annotations (labelâĹŠvalue pairs); this is disabled for the class *Peak* and all its derivatives to keep memory usage at a minimum.

generally quite large, which can cause problems with the inefficient memory usage of Java programming objects. To circumvent these limitations, the class *PeakData* was introduced, which has memory-efficient arrays for: scan numbers, retention times, masses, intensities, and measurement-ids. The classes *Spectrum*, *MassChromatogram*, and *BackgroundIon* use this class to store all the data, minimizing the memory requirements. To ease interaction with *PeakData* from the end-programmer, there are two helper classes *Centroid*, and *Profile* to easily interact with stored data.

Next to the data classes, a series of meta data "header" classes are provided to describe the data sets. In addition a comprehensive set of I/O routines is provided for data loading and writing of mass spectrometry data in the major open file formats mzData, mzXML, mzML, next to PeakML. As each of these formats follows roughly the same format in terms of meta-information, they can be loaded into the common header classes. The class *Header* is the entry point, binding the following classes: *MeasurementInfo* with measurement-specific information, such as associated files, uniquely identifiable by a measurement-id also stored in *IPeak*; *SetInfo* that combines multiple measurements into a set; *SampleInfo* describes sample specific information, stored in annotations; *ApplicationInfo* describes software application-specific information that was used to produce the data, such as software name, version, parameters, etc. Additionally, 1-to-1 mappings are provided to the file access libraries of Waters Corp. and Thermo Fisher Scientific Inc., both of which are only accessible on Microsoft Windows platforms (due to the implementation chosen by these manufacturers).

The raw output file from a mass spectrometry run contains large amounts of meta- and status-information, which is rarely (if at all) preserved during the transformation process to an open standard such as mzXML. For further interpretation of the data, this meta-information can play a key role, providing insight into the quality of each measurement. For this reason, we additionally supply the tool ThermoLogViewer, which allows the user to load multiple RAW-files and compare their logs (see online documentation at `http://mzmatch.sourceforge.net` for more information).

The PeakML library allows for quick access to the content of data files. As a reference, we used the data files listed in Table 2.1. On a Dell Optiplex 780 desktop computer with an Intel Core2 Duo Q9550 CPU and 8 GB of RAM, loading PeakML files containing features for single LC–MS runs required 2 s. The initial loading and displaying of a massive PeakML file containing 23926 peak sets required 13 s. Once loaded into memory, all extracted ion chromatograms can be accessed instantly.

2.1.3 The mzmatch.R R package

The PeakML file format enables the uniform exchange of intermediate and result data between analysis software from different manufacturers and groups. This is important because each piece of software has its own unique strengths and weaknesses. Cherry-picking of components enables researchers to construct a data analysis pipeline specifically suited to their needs. Use of PeakML can enable such flexible pipelines, which we illustrate here by integrating PeakML with the R-package XCMS (Smith *et al.*, 2006). This package has excellent support for data processing, statistics, and graph visualization. However, R is primarily targeted at programmers, potentially locking nonexpert users out from further data analysis. Moreover, it is difficult to visualize the extracted peaks such that one can browse through them and select the peaks of interest (e.g., the getEIC routines of XCMS are not straightforward to apply). In contrast, this functionality is easily implemented in the framework with the PeakML file format, as shown with the PeakML Viewer. The R-package mzmatch.R extends XCMS with functionality for storing data in PeakML files and vice versa, such that these tools can be connected (see online documentation at `http://mzmatch.sourceforge.net` for more information). Because PeakML includes system-defined annotations (e.g., "identifications": a comma-separated list of database id's; "relation.id" and "relation.ship": identifiers for derivative peaks and their relationships), visualizations beyond the current capabilities of XCMS are enabled.

2.2 Applications

2.2.1 mzMatch

The availability of the PeakML file format makes it possible to split the components of a processing pipeline into small tools (peak extraction, alignment, noise filtering, etc., described in refs Scheltema *et al.* (2008) and Scheltema *et al.* (2009)) that can easily be connected into various configurations. This principle has been applied to the implementation of the data analysis pipeline mzMatch (Figure 2.5; an extensive step-by-step tutorial on the example pipeline is available at `http://mzmatch.sourceforge.net/tutorial.mzmatch.r.advanced.html`). The tools included are mass chromatogram extraction, matching (called "grouping" in XCMS), derivative detection (Scheltema *et al.*, 2009), noise filtering, normalization, and alignment (see online documentation at `http://mzmatch.sourceforge.net/mzmatch/index.html` and `http://mzmatch.sourceforge.net/mzmatch.R/00Index.html`). For example,

the `RelatedPeaks` (Scheltema *et al.*, 2009) tool is very effective in gathering all features caused by a single analyte (including isomers) and annotating them accordingly. This means the features are not removed from the data, but only tagged, allowing for later inspection by the analyst. It is the responsibility of the analyst to validate the identity of the peaks with additional, orthogonal biochemical techniques or internal standards.

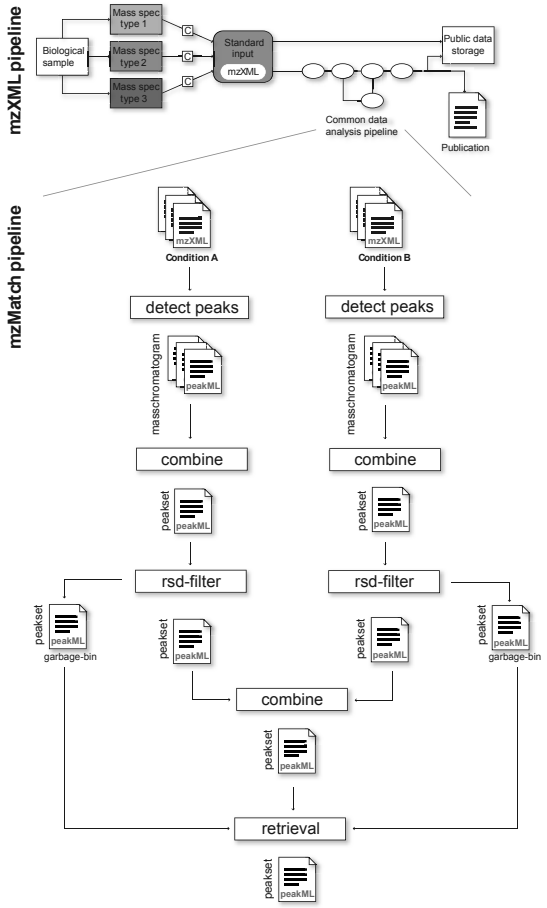


Figure 2.5: An example of an mzMatch pipeline. The mzXML/mzData/mzML standards changed the data analysis landscape by providing a common input format but do not provide functionality for breaking up a data analysis pipeline into small interchangeable components. The mzXML paper (Pedrioli *et al.*, 2004) proposed the pipeline shown at the top but did not go into further detail about the setup of a common data analysis pipeline. The mzMatch pipeline makes effective use of the PeakML file format for defining small components in a data analysis pipeline. The data in each file can be picked up by any tool; e.g., in this small example a noise filter could have been introduced right after the `detect peaks` tool, without breaking the pipeline.

In addition, all filter tools discarding signals from the data set, such as the CoDA-DW (Windig, 2004) noise filter *mzmatch.filter.NoiseFilter*, also export the discarded signals next to the result data set. Such behavior provides traceability for the performance of each tool, as the user can easily verify whether the operation had the desired effect. It also offers the potential for retrieving signals of interest lost in one of

the filter steps which cannot easily be achieved with other data processing packages. The RSD filter (Shah *et al.*, 2000), *mzmatch.filter.RSDFilter*, removes signals that were irreproducible in biological and/or technical replicates. However, when multiple experimental conditions are used, it can happen that the behavior for signals in one condition falls outside the specified range of reproducibility, but not in the other(s). Such signals can then be recovered in all experiments with the recycle bin recovery tool, *mzmatch.util.Recovery*.

As described previously, the output of each tool can readily be used in either XCMS or the mzMatch/PeakML Viewer. Each tool is command-line-based such that settings can be passed to the tools as command-line options (e.g., "-i" for the input file(s)). The mzMatch toolbox is designed in such a way that documentation for each tool can automatically be generated. In addition, tools can also be automatically exposed in a programming language (such as the R environment) as functions or a pipeline workflow environment such as Taverna (Kuhn *et al.*, 2010). This has been done to extend the mzmatch.R library with the mzMatch tools.

2.2.2 PeakML viewer

A user interface application called PeakML Viewer (Figure 2.6) enables rapid visualization, inspection, and manipulation of the contents of a PeakML file (e.g., manual selection and/or export of peaks of interest). After a PeakML file is loaded, the "entry" view gives an overview of all the entries with the retention time, mass, and intensity. The ordering from the original file is kept intact, making the results from sorting tools such as *mzmatch.ipeak.sort.RelatedPeaks* (Scheltema *et al.*, 2009) accessible. An entry is highlighted in bold when it has been matched to a compound from a database with *mzmatch.ipeak.util.Identify* (determined by verifying whether the entry contains the system-defined annotation *identification*). By clicking on an entry, the associated traces will be displayed in the graph view and the identifications in the "identification" view (including the mass deviation in ppm and the putatively assigned molecular structure when available). There is an additional tab *derivatives*, which shows all the clustered related peaks with their identification as determined by *mzmatch.ipeak.sort.RelatedPeaks* (stored in the system defined annotations *relation.id* and *relation.ship*). The "filter" view allows the user to perform simple operations on the data (sorting and filtering), for zooming in on the entries of interest. The "trend" view gives an overview of the intensity levels for each entry (which in the case of a peakset can for example consist of multiple mass chromatograms). All the peaks belonging to the same set are grouped together in this plot, and the mean,

minimum, and maximum values are displayed. The "sets" view shows all the measurements used and how they are organized in sets. With the check boxes, all the peaks from a set or peaks individually can be switched on or off (i.e., not displayed). The "annotations" view gives an overview of all the annotations that are available for the current entry.

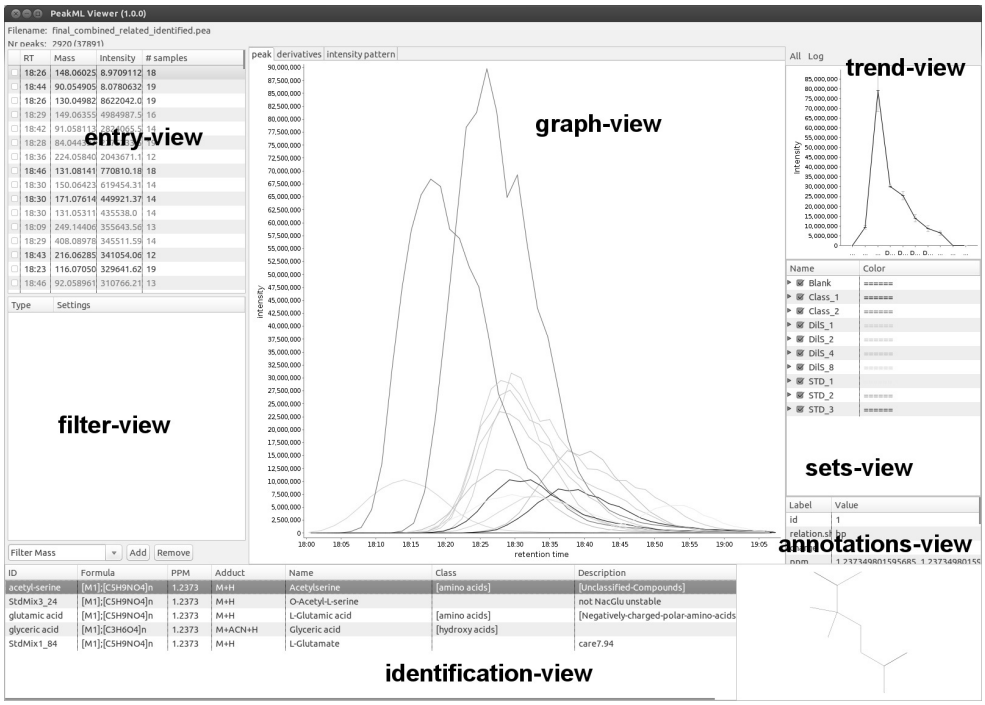


Figure 2.6: Screenshot of the PeakML Viewer. The viewer enables the user to load PeakML files (generated at any point in the analysis pipeline, providing full control and verifiability of the data processing steps), visualize, and browse through its contents. The interface is divided into seven views, providing all the information stored in the file. The most important are the "entry" view, giving an overview of all the IPeak entries stored in the file, and the "graph" view, giving a visual representation of the data stored for the entry. By using the keyboard (arrow up/down and spacebar for selection), the user can rapidly verify the contents and select peaks of interest.

2.3 Discussion

The PeakML file format enables research groups to transcend the monolithic development model of mass spectrometry data analysis software and start building flexible, modular application pipelines. The benefits include (i) increased verifiability of the performance of individual analysis steps, (ii) an easy "rewind" option to roll back to intermediate steps in the analysis process, and (iii) the straightforward use of analytical components from alternative pipelines originally not intended by the software authors. Moreover, tool developers can have a much broader user group for their software, because its components can be more easily recombined to suit the needs of different researchers. With the PeakML and mzmatch.R libraries, a first successful integration between data analysis environments created by different groups has been demonstrated. Of course, PeakML still has limitations. For example, both PeakML and the mzMatch toolbox have been developed mostly for metabolomics experiments (Kol *et al.*, 2010; t'Kindt *et al.*, 2010a,b), but their functionality for proteomics experiments is in its infancy. Additionally, adding an indexing mechanism to speed up searches in large PeakML files is desirable. We would like to invite other groups to join in this open source development at `http://mzmatch.sourceforge.net` to enable the larger mass spectrometry community to share the best data and tools notwithstanding large variation in research aims.

PeakML/mzMatch highlights

- Fully documented, complete, and platform-independent mass spectrometry specific API complete with programming examples.
- Defines the PeakML file format, offering functionality to store and share processed data for further processing and to publish verifiable results.
- Supports the major file formats mzData, mzXML, mzML; provides a 1-to-1 mapping to the file access libraries of Waters Corp. and Thermo Fisher Scientific Inc.
- Plays well with others by offering the potential to integrate with other software, as illustrated by the integration with XCMS.
- Integrated chemistry (e.g., molecular formulas, mass conversio n, periodic table), math (e.g., statistics, wavelet transform, function fitting, and loess and

Savitzky-Golay), and visualization (JFreeChart and SWT for user interface applications) routines.

- A set of small and agile tools (e.g., mass chromatogram extraction, combining, noise filtering, normalization) performing defined operations on the data.

Acknowledgements

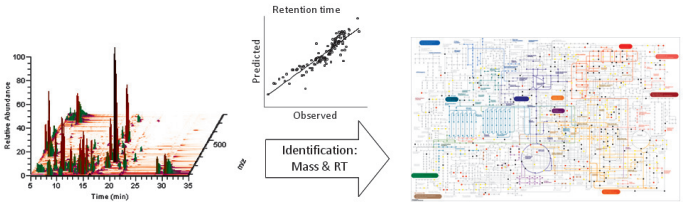
R.A.S. and A.J. contributed equally to this work. The authors gratefully acknowledge the efforts and contributions of Elena Merlo and George Byelas (University of Groningen, The Netherlands), Saskia Decuypere and Ruben t'Kindt (Institute of Tropical Medicine, Belgium), Martijn Dijkstra and Marcel de Vries (University Medical Centre Groningen, The Netherlands), David Wildridge, Jana Anderson, Isabel May-Vincent, Darren Creek, Karl Burgess, and Michael Barrett (University of Glasgow, UK), Fabien Jourdan (INRA, France), and Anas Kamleh and David Watson (Strathclyde University, UK) for providing data, testing the software, and suggesting improvements and additional features. R.B. and R.C.J. devised and supervised the project. R.A.S. designed and implemented the software architecture and the PeakML file format. A.J. designed and implemented the XCMS integration and R-tools accompanying the software. R.S., A.J., M.A.S., and R.B. wrote the manuscript. R.B. is supported by a Netherlands Organisation for Scientific Research (NWO) Vidi fellowship. R.A.S. is supported by a NWO Vici grant to RCJ. M.A.S. is supported by NWO Rubicon (825.09.008) and The Netherlands Bioinformatics Center. R.B. and M.A.S. are supported by The Netherlands Proteomics Center (NPC-GM WP 1.1).

Chapter 3

Toward global metabolomics analysis with Hydrophilic Interaction Liquid Chromatography–mass spectrometry: improved metabolite identification by retention time prediction

Darren J. Creek[1,2], Andris Jankevics[3,4], Rainer Breitling[3,4], David G. Watson[5], Michael P. Barett[1] and Karl E. V. Burgess[1]

Anal. Chem. 83(22):8703-10.



1 Institute of Infection, Immunity and Inflammation, Wellcome Trust Centre for Molecular Parasitology, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, UK
2 Department of Biochemistry and Molecular Biology, Bio21 Molecular Science and Biotechnology Institute, University of Melbourne, Melbourne, Australia
3 Institute of Molecular, Cell and Systems Biology, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, United Kingdom
4 Groningen Bioinformatics Centre, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Groningen, The Netherlands
5 Strathclyde Institute of Pharmacy and Biomedical Sciences, University of Strathclyde, Glasgow, United Kingdom

abstract

Metabolomics is an emerging field of postgenomic biology concerned with comprehensive analysis of small molecules in biological systems. However, difficulties associated with the identification of detected metabolites currently limit its application. Here we demonstrate that a retention time prediction model can improve metabolite identification on a hydrophilic interaction chromatography (HILIC)–high-resolution mass spectrometry metabolomics platform. A quantitative structure retention relationship (QSRR) model, incorporating six physicochemical variables in a multiple-linear regression based on 120 authentic standard metabolites, shows good predictive ability for retention times of a range of metabolites (cross-validated $R^2 = 0.82$ and mean squared error = 0.14). The predicted retention times improved metabolite identification by removing 40% of the false identifications that occurred with identification by accurate mass alone. The importance of this procedure was demonstrated by putative identification of 690 metabolites in extracts of the protozoan parasite Trypanosoma brucei*, thus allowing identified metabolites to be mapped onto an organism-wide metabolic network, providing opportunities for future studies of cellular metabolism from a global systems biology perspective.*

Metabolomics is a rapidly growing field of postgenomic biology, aiming to comprehensively characterize the small molecules in biological systems. While genomics, transcriptomics, and proteomics enable untargeted investigation of cells, tissues, and organisms, metabolite analysis potentially offers the most direct measure of the phenotypic state of a biological system (Breitling *et al.*, 2008; Kell, 2006). Our analytical ability to measure the global metabolome is currently limited by the chemical diversity of biological metabolites, and metabolomics studies generally follow one of two compromised approaches: targeted or untargeted (Dunn *et al.*, 2011b; Scalbert *et al.*, 2009). Targeted metabolic profiling approaches require a priori knowledge of the metabolites of interest, and the analytical platform is optimized to detect these metabolites quantitatively (Lu *et al.*, 2008). This approach has been successfully applied to the investigation of specific metabolic pathways (Olszewski *et al.*, 2010) but does not achieve a global coverage. The alternative untargeted approach involves statistical analysis of all signals from an analytical platform and subsequent identification of the most significant metabolites (De Vos *et al.*, 2007). This approach is often employed in hypothesis-generating studies such as biomarker discovery (Sreekumar *et al.*, 2009), but exhaustive metabolite identification is generally not the goal. The scope and confidence of metabolite identification is a major bottleneck in the

interpretation of metabolomics data, and improvements in metabolite identification are urgently required (Dunn *et al.*, 2011b). Comprehensive metabolite identification would allow the construction of biologically meaningful metabolic networks from metabolomics data and allow metabolic biochemistry to be investigated from a systems biology perspective (Breitling *et al.*, 2006b).

Liquid chromatography coupled to mass spectrometry (LC–MS) is becoming increasingly popular for metabolomics analysis (Cubbon *et al.*, 2010; Scalbert *et al.*, 2009). Reversed phase LC–MS is well established for small molecule analysis and is particularly useful for separation of hydrophobic metabolites as it involves elution of analytes by an aqueous-based mobile phase (often with a gradient of increasing organic content) from a nonpolar stationary phase. In contrast, hydrophilic interaction chromatography (HILIC) utilizes a gradient of increasing aqueous content to elute analytes from a hydrophilic stationary phase. While reversed phase methods are commonly used for LC–MS based metabolomics, HILIC is becoming an attractive alternative or complementary approach, due to the ability to separate hydrophilic metabolites (Cubbon *et al.*, 2010; Dunn *et al.*, 2011b; Scalbert *et al.*, 2009). The electrospray (ESI) ionization source commonly applied to interface LC with MS generally provides good sensitivity and a high proportion of molecular ions for detection. In addition to molecular ions, many additional ion signals are also acquired for each metabolite, such as in-source fragments, adducts, and multiply charged species, which often complicate interpretation of LC–MS data and may lead to false metabolite identifications (Brown *et al.*, 2009). Recent advances in ultrahigh resolution mass detectors often allow direct assignment of chemical formulas by accurate mass detection within 1 ppm (Breitling *et al.*, 2006b; Moco *et al.*, 2007). Accurate mass detection opens the door for untargeted metabolite identification but is not always sufficient for unambiguous identification, primarily due to the complexity of LC–MS data and the large number of isomeric compounds present in biological systems. Orthogonal information is therefore required to confirm identification, with LC retention time (RT) and MS/MS (or $MS_{(n)}$) fragmentation being the most commonly used, and easily obtained, orthogonal data (Scalbert *et al.*, 2009). Metabolite identification ideally requires analysis of authentic standards, so that retention time and/or fragmentation patterns can be compared (Brown *et al.*, 2009). However, it is not practical for individual laboratories to purchase and analyze authentic standards for every possible metabolite nor are all metabolites commercially available. Public MS/MS fragmentation databases such as Massbank (`http://www.massbank.jp`) (Horai *et al.*, 2010) and Metlin (`http://metlin.scripps.edu/`) (Smith *et al.*, 2005) offer some help toward resolution of the problem, but although fragment patterns are consistent for a

given fragmentation type (e.g., collision induced dissociation (CID), electron capture dissociation (ECD)), fragment intensities are often instrument-specific, and these databases are limited to metabolites that have commercially available standards. Fragmentation prediction (such as `http://msbi.ipb-halle.de/MetFrag/`) may offer a useful alternative where standard MS/MS data is not available, but current prediction algorithms are not accurate for all metabolites (Wolf *et al.*, 2010). Retention time also provides useful orthogonal information for metabolite identification where authentic standards can be analyzed on the same platform (Brown *et al.*, 2009); however, to our knowledge no attempt has been made to store RT information in public databases for LC platforms, due to the significant variability between platforms. Even temporal variability between batches on the same platform can occur due to changes in column chemistry (aging or batch variation), mobile phase composition, and temperature.

Quantitative structure retention relationships (QSRR) allow prediction of HPLC retention time based on the physicochemical nature of the analyte–column interactions that determine retention (Kaliszan, 2007). This form of quantitative structure–property relationship requires experimental measurement of retention times for a training set of authentic standards and determination of chemical descriptors, which can be calculated from compound structures computationally. This allows development of a model with which retention times for a large database of metabolites can be predicted based on their calculated physicochemical properties. QSRR has been successfully applied to HPLC for specific classes of compounds, such as peptides (Kaliszan *et al.*, 2005) and steroids (Salo *et al.*, 1996); however, the application to a large, chemically diverse group of metabolites is somewhat more ambitious. Nevertheless, the prediction of a retention time window, in conjunction with accurate mass, will allow greatly improved annotation of metabolite identities in large-scale metabolomics studies (Kind and Fiehn, 2010), as was earlier shown with a retention prediction model for capillary electrophoresis–mass spectrometry (CE–MS) (Sugimoto *et al.*, 2010).

This article describes the development of a validated QSRR model for HILIC chromatography and application of the model to a large metabolite database. A template is provided to recalculate structure–retention relationships for specific analytical platforms to allow application of this methodology in other laboratories. Furthermore, we demonstrate the benefit of incorporating predicted retention times into the metabolite identification procedure for both metabolite standards and biological samples.

3.1 Experimental section

3.1.1 Preparation of authentic standards

Stock solutions were prepared for each one of a set of 127 authentic standard compounds at 10 or 100 mM in either Milli-Q water, ethanol, or 50% ethanol/water, depending on solubility, or in the case of less soluble compounds at lower concentration or in alternative solvents, as detailed in Supporting Information S1. Stock solutions were combined into groups of approximately 15 compounds at 100 μM for determination of retention times, ensuring that there were no isobaric compounds within each group and no combination of compounds where likely MS fragments would be isobaric with other compounds and thus hinder assignment of retention time to an individual compound. In the few cases where assignment of retention time was ambiguous, the individual compound was subsequently analyzed separately. A comprehensive standards solution was also prepared, containing all 127 authentic standards at 100 μM concentration. This comprehensive standards solution was further diluted with 80% acetonitrile/water to 10, 1, and 0.1 μM solutions for analysis.

3.1.2 LC–MS method

The LC separation was performed using hydrophilic interaction chromatography with a ZIC-HILIC 150 mm × 4.6 mm, 5 μm column (Merck Sequant), operated by a Dionex UltiMate liquid chromatography system (Dionex, Camberley, Surrey), and coupled to a FAMOS autosampler. The LC mobile phase was a linear gradient from 80% B to 20% B over 30 min, followed by an 8 min wash with 5% B, and 8 min re-equilibration with 80% B, where solvent B is 0.08% formic acid in acetonitrile and solvent A is 0.1% formic acid in water. The flow rate was 300 μL/min, column temperature 20 °C, injection volume 10 μL, and samples were maintained at 4 °C. The mass spectrometry was performed using an Orbitrap Exactive (Thermo Fisher Scientific, Hemel Hempstead, U.K.) with a HESI 2 probe. The spectrometer was operated in polarity switching mode, with the following settings: resolution 50000, AGC 1 × 106, m/z range 70–1400, sheath gas 40, auxiliary gas 20, sweep gas 1, probe temperature 275 °C, and capillary temperature 250 °C. For positive mode ionization: source voltage +4 kV, capillary voltage +50 V, tube voltage +70 V, skimmer voltage +20 V. For negative mode ionization: source voltage –3.5 kV, capillary voltage –50 V, tube voltage –70 V, skimmer voltage –20 V. Mass calibration was performed for each polarity immediately before each analysis batch, within 48 h for all samples. The calibration mass range was extended to cover small metabolites by inclusion of

low-mass contaminants with the standard Thermo calmix masses (below m/z 1400), acetonitrile dimer for positive ion electrospray ionization (PIESI) mode (m/z 83.0604) and C3H5O3 for negative ion electrospray ionization (NIESI) mode (m/z 89.0244). To enhance calibration stability, lock-mass correction was also applied to each analytical run using these ubiquitous low-mass contaminants.

3.1.3 Database generation

A comprehensive metabolite database, containing 41623 potential metabolites (Supporting Information S2), was constructed by combining metabolite entries from the publicly available online metabolite databases KEGG (`http://www.genome.jp/kegg/`) (Kanehisa *et al.*, 2010), MetaCyc (`http://www.metacyc.org`) (Caspi *et al.*, 2010), HMDB (`http://www.hmdb.ca`) (Wishart *et al.*, 2009), and Lipidmaps (`http://www.lipidmaps.org`) (Sud *et al.*, 2007) with an internally generated database of dipeptides, tripeptides, and tetrapeptides of proteinogenic amino acids. For each metabolite, the database contains the exact mass, chemical formula, name, and where available, meta-information including KEGG or MetaCyc pathways, identifiers, and SMILES strings (Weininger, 1988). Missing SMILES strings in source databases were imported from Chemspider (`http://www.chemspider.com`) where possible. InChi Keys were imported from the Chemical Translation Service (`http://cts.fiehnlab.ucdavis.edu`) (Wohlgemuth *et al.*, 2010), and available identifiers were used to minimize redundancy by merging synonymous entries from multiple data sources. Where multiple isomers exist for a given formula, the database was sorted to give preference to the most biologically "likely" metabolites based on (1) genome-annotation for the organism of study (e.g., TrypanoCyc (Chukualim *et al.*, 2008)), (2) metabolites of central metabolic pathways in KEGG (amino acid, carbohydrate, energy, nucleotide, and lipid metabolism), (3) number of databases containing each metabolite (KEGG, MetaCyc, HMDB, Lipidmaps). The final sorting is based on metabolite ID numbers in the source databases, as the historical development of these databases that has resulted in a trend toward assignment of lower ID numbers to common metabolites involved in primary metabolism and higher ID numbers to more unusual metabolites or xenobiotics.

Where SMILES strings were available, Jchem for Excel (Jchem For Excel 5.3.1, 2010, Chemaxon, `http://www.chemaxon.com`) was used to calculate a number of physicochemical properties for the metabolites in the database (Supporting Information S2). pH-dependent parameters (log D and charge) were calculated throughout the apparent pH range of the mobile phase, from 2.65 (aqueous phase) to 4.5 (organic phase), and 3.5 was chosen for all further calculations because it provided the best fit

for the model. Positive and negative charges were calculated for each pH by addition of the formal charge of the molecule to the relative charge on each ionizable group according to predicted pKa values and the Henderson–Hasselbalch equation (eq. 3.1 and 3.2).

$$Neg = NC + \sum_{x=1}^{N_a}\left(\frac{10^{(pH-pK_a(x))}}{1+10^{(pH-pK_a(x))}}\right) \tag{3.1}$$

NC = formal negative charge, N_a = number of acidic functional groups

$$Pos = PC + \sum_{x=1}^{N_b}\left(\frac{10^{(pK_a(x)-pH)}}{1+10^{(pK_a(x)-pH)}}\right) \tag{3.2}$$

PC = formal positive charge and N_b = number of basic functional groups.

3.1.4 QSRR calculations

Experimental retention times for 120 metabolite standards (with molecular weights ranging from 70 to 400) were determined by LC–MS analysis with Xcalibur and ToxID (Thermo Fisher Scientific), were converted to retention factors (RF) (eq. 3.3), and entered into a multiple linear regression (MLR) model with calculated chemical descriptors (from Jchem for Excel) as variables (Supporting Information S2). Chemical descriptors were considered for inclusion only if they could be rapidly calculated from SMILES strings by freely available software and their potential to influence retention time could be readily explained in physicochemical terms.

$$RF = \frac{(RT - column\ void\ time)}{column\ void\ time} \tag{3.3}$$

The "leaps" package (Lumley and Miller, 2009) in the statistical software R (R Core Team, 2012), was used for selection of the most descriptive variables by an exhaustive search of 11 physicochemical properties (Supporting Information S3). The optimal regression model was assessed by Mallow's Cp statistic (Mallows, 1973). The selected model was evaluated by a 10-fold cross-validation, and the predictive power was estimated by validated adjusted R^2 and the mean squared error (MSE) of prediction.

3.1.5 Metabolomics sample preparation

For an illustrative test case, metabolites were extracted from bloodstream-form *Trypanosoma brucei brucei* (strain 427), cultured *in vitro* with HMI-9 medium and 10% fetal calf serum (Hirumi and Hirumi, 1989), to a density of 2 × 106 cells/mL. Appropriate volumes of cell culture were taken to yield 5 × 107, 1 × 108, and 2 × 108 total cells per sample. Cells were concentrated by centrifugation at 1000g for 10 min at 37 °C, the cell pellet was resuspended in 1 mL of supernatant and transferred to a 1.5 mL tube, and the final cell pellet was obtained by further centrifugation at 3000g for 5 min and complete removal of supernatant. Metabolites were extracted from the cell pellet by addition of 200 μL of chloroform/methanol/water (1:3:1) with vigorous mixing for 1 h at 4 °C (t'Kindt *et al.*, 2010b).

Precipitated proteins and cellular debris were removed by centrifugation at 13000g for 5 min, and the supernatant was kept at –80 °C until LC–MS analysis within 2 weeks. Additional control samples included cell-free growth medium and extraction solvent blanks. All samples were prepared in triplicate from a single parasite culture.

3.1.6 Metabolomics data processing

Raw LC–MS data were processed with a combination of XCMS Centwave for peak picking (`http://metlin.scripps.edu/xcms`) (Smith *et al.*, 2006) and mzMatch for alignment and annotation of related peaks (`http://mzmatch.sourceforge.net`) (Scheltema *et al.*, 2011). Metabolite identification was performed by matching masses and retention times to the database with a mass accuracy window of 3 ppm (if two formulas were within 3 ppm the closest match was taken) and RT window of 35% (by in-house VBA scripts). Additional automated noise and MS artifact filtering procedures were applied to remove peak sets that contained (1) peaks that were present at equal or higher abundance in the blank solvent samples, (2) all peaks lower than the intensity threshold (10000), (3) shoulder peaks or duplicate peaks within the same mass (3 ppm) and retention time (0.2 min) window, (4) common MS artifacts accord-

ing to the table in Supporting Information S4 and S5 with irreproducible intensities (relative standard deviation > 0.5) across replicate samples (biological study only) (Scheltema *et al.*, 2009).

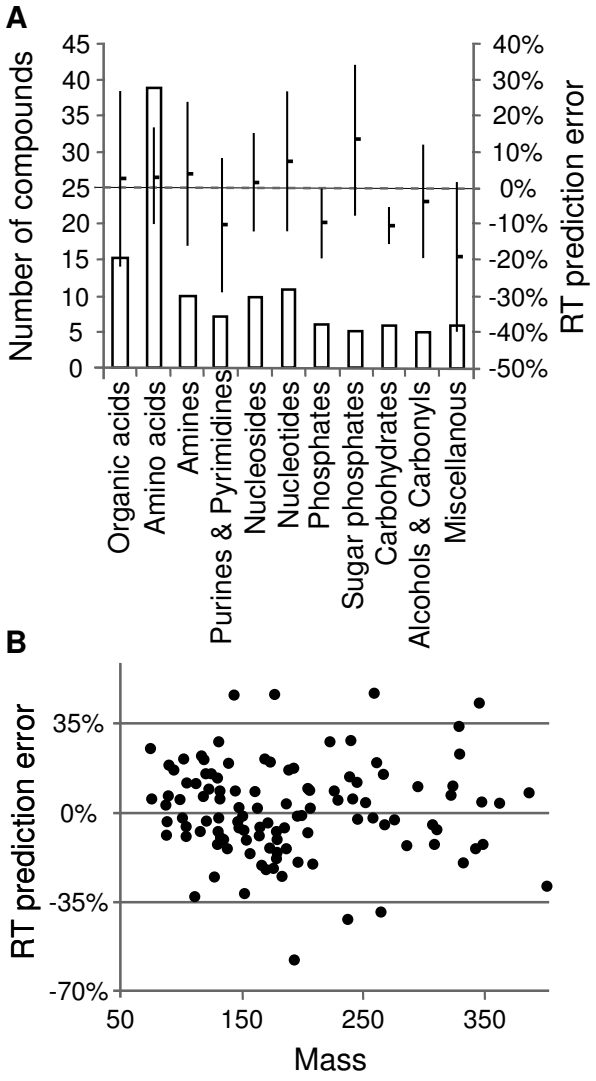


Figure 3.1: A) Number of authentic standard metabolites in each metabolite class (columns) and retention time prediction accuracy for each class (mean marked by small black squares; standard deviation marked by error bars). (B) Distribution of retention time prediction errors for authentic standard metabolites according to metabolite mass.

3.2 Results and discussion

3.2.1 LC–MS method

The HILIC–Orbitrap LC–MS method proved successful for detection of many metabolites (Kamleh *et al.*, 2008, 2009b). While it is generally accepted that a truly comprehensive metabolomics approach is not possible on a single platform (Garcia *et al.*, 2008; Moco *et al.*, 2007), the HILIC-Exactive approach, with polarity switching, is shown here to detect authentic standards from a wide range of metabolite classes (Figure 3.1 and Supporting Information S5).

In addition to the identification of authentic metabolite standards, this analytical method produces thousands of additional peaks, many of which can be putatively annotated as metabolites based on exact mass (Supporting Information S6). However, closer inspection of these peaks reveals many retention times that are incongruous with the expected retention time of these metabolites according to the basic principles of hydrophilic interaction chromatography on a ZIC-HILIC column (`www.sequant.com/hilic`), and in some cases, these annotations were confirmed to be false identifications by comparison with authentic standards (where available). The false identifications may arise from a number of sources including MS artifacts (e.g., fragments, adducts, isotopes), chromatographic issues (e.g., peak shoulders, poor retention, peak-picking errors), and noise (e.g., contaminants and MS signal processing artifacts). In order to systematically remove these misleading peaks from a metabolomics data set, a retention time prediction model was developed to support identification based on both accurate mass and retention time.

3.2.2 QSRR modeling and validation

The retention time prediction model was developed using the 120 authentic standards with experimentally determined retention times on this analytical platform (Supporting Information S5). The optimal QSRR model (eq. 3.2.2), as determined by the MLR model selection (Supporting Information S3) has an adjusted R^2 of 0.82 (Figure 3.2), has a residual standard error of 0.35, and includes 6 calculated physicochemical properties as variables: log D = calculated octanol–water partition coefficient at pH 3.5, Neg = negative charge at pH 3.5, Pos = positive charge at pH 3.5, Rot = number of rotatable bonds, Phos = number of phosphate groups, and (HBD/MW) = number of hydrogen bond donors divided by molecular weight.

$$log(RF) = k_1(logD) + k_2(Neg) + k_3(Pos) + k_4(Rot) + k_5(Phos) + k_6(HBD/MW) + constant \quad (3.4)$$

Regression coefficients: k_1 = –0.2449, p = 4.9×10^{-13}; k_2 = 0.2397, p = 0.0035; k_3= 0.1967, p = 0.01; k_4 = –0.0523, p = 0.041; k_5 = 0.4599, p = 4.5×10^{-5}; k_6 = 19.8424, p = 0.0037; constant = –0.8747, p = 5.2×10^{-10}.

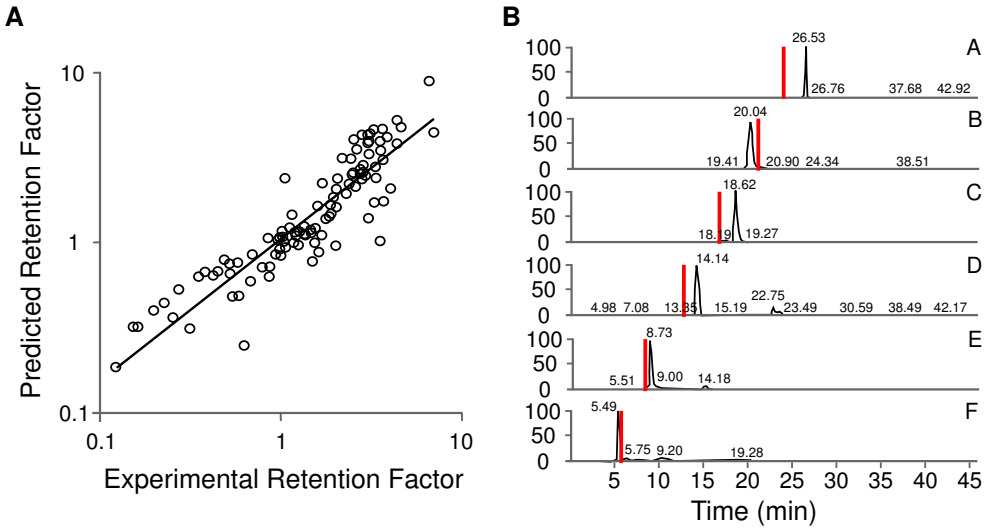


Figure 3.2: (A) Experimental vs predicted retention factors for authentic standards in the QSRR model. (B) Extracted ion chromatograms and predicted retention times (red bars) for (A) putrescine, (B) d-ribose 5-phosphate, (C) *sn*-glycero 3-phosphocholine, (D) *L*-proline, (E) xanthine, (F) lipoate.

The most predictive variable is log D (p = 4.9×10^{-13}), the octanol–water partition coefficient as calculated by Jchem at pH 3.5. This finding confirms the general mechanism of hydrophilic interaction chromatography, which involves partitioning of the analyte between the organic mobile phase and aqueous stationary phase. The five additional variables are also involved in phase partitioning and/or hydrophilic or ionic analyte–column interactions, thus supporting the validity of the model from a physicochemical perspective (Kaliszan, 2007).

A 10-fold cross-validation of the model showed good predictive ability (MSE = 0.14, adjusted $R^2 = 0.82$, and residual standard error = 0.35). Perhaps most important for application to the identification of metabolites in metabolomics studies is the retention time window that can be predicted with confidence for each possible metabolite (Kind and Fiehn, 2010). In this regard, we observed that the true retention times for 93% of metabolite standards were within 35% of the predicted retention times (for example, ±1.75 min for a compound eluting at 5 min or ±7 min for a compound eluting at 20 min). Significant outliers were primarily due to inherent errors associated with the high dependence of the model on calculated values, log D and pK_a. The largest error was associated with ascorbate, which has a unique ionizable group, with a miscalculated pK_a of 0.5 (i.e., predominantly charged at mobile phase pH 3.5), yet an experimental pK_a of 4 (i.e., predominantly uncharged at pH 3.5) (Khan and Martell, 1967). Replacement of the calculated log D value (–4) with the calculated log P value (–1.3) (to represent the uncharged species) resulted in an accurate RT prediction (<15% error). The overestimate of the retention time for AMP can also be explained by underestimation of the log D and log P values in Jchem, when compared with other calculated log P values obtained from HMDB (`http://www.hmdb.ca`) (Wishart *et al.*, 2009). The error associated with citric acid was most likely due to the poor chromatographic behavior of this polyacidic metabolite, which exhibited a very wide peak (>5 min) and irregular peak shape.

It should be noted that all compounds included in the QSRR model had molecular weights below 400. Authentic standards for 7 common larger metabolites (above 400 Da) were analyzed; however, these retention times were poorly predicted by the model, most likely due to errors associated with predicted log D. Therefore, the model should only be applied to compounds with molecular weights below 400 (which, by definition, accounts for the vast majority of biological metabolites aside from lipids and triphosphates, which are not ideally suited to this analytical platform (Kamleh *et al.*, 2008)). Despite the inability of the QSRR model to accurately predict retention times for larger molecules, a feature of HILIC chromatography is the early elution of hydrophobic compounds, including lipids (Kamleh *et al.*, 2008), thus allowing a class-based retention time prediction for most large metabolites based on predicted log P (log $P > 0$) and/or classification in the lipidmaps database (Sud *et al.*, 2007). The combination of the QSRR model and the class-based prediction for large hydrophobic metabolites resulted in assignment of predicted retention times for over 92% of all compounds in the database of potential metabolites (excluding peptides).

3.2.3 Application to untargeted metabolite analysis

LC–MS data preprocessing techniques for metabolomics are gradually improving as the number of open-source applications for this purpose expands. XCMS (Smith *et al.*, 2006), MZmine2 (Pluskal *et al.*, 2010b), Maven (Melamud *et al.*, 2010), and mzMatch (Scheltema *et al.*, 2011) all provide peak-picking capability and some capacity for peak filtering and annotation. However, accurate metabolite identification remains a major bottleneck in untargeted metabolomics studies (Matsuda *et al.*, 2009). Data produced by these applications contain many peaks that are not molecular ions of biological metabolites (Brown *et al.*, 2009; Scheltema *et al.*, 2009), resulting in numerous false identifications when accurate mass alone is used to identify peaks and new software to improve annotation of these ions has recently been developed (Brown *et al.*, 2011; Scheltema *et al.*, 2011). Retention time prediction offers a new approach to filtering metabolite identities, using information which is routinely collected but not directly related to metabolite mass.

To demonstrate the impact of retention time prediction on untargeted metabolite identification, the 4 concentrations of comprehensive standard solutions were analyzed, revealing 20150 peak sets (PIESI mode only). Matching the accurate mass of these peaks to the metabolite database (with a 3 ppm window) resulted in 3133 putative identifications, suggesting that a large number of false identifications were observed from these mixtures of 127 metabolite standards, even taking into consideration that few biochemical standards are purchased at 100% purity. However, an additional filter that requires the observed retention time of putative metabolites to be within 35% of the predicted retention time, removed 40% (1314) of the putative identifications (Supporting Information S6). Only 6 of the authentic metabolite standards were lost by this filtering process, confirming the effectiveness of retention time prediction in metabolite identification procedures.

When used in combination with other routinely used data-dependent noise and artifact filters (see Experimental Section), the total number of putatively identified peaks is reduced to 627, an 80% reduction in data (from 3133 putative identifications in the unfiltered set), yet retaining 91% of the detected authentic standards (Supporting Information S4).

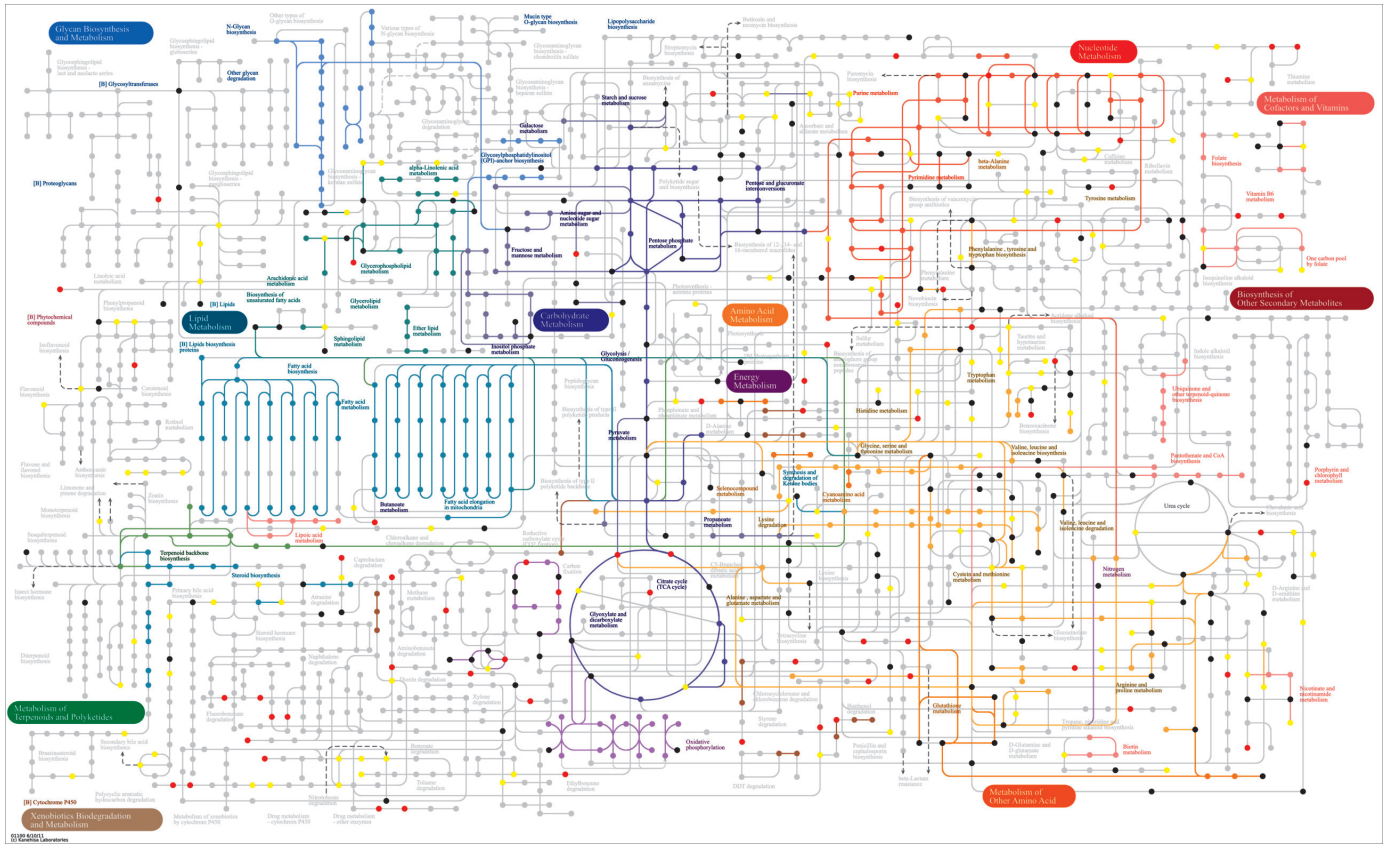


Figure 3.3: *T. brucei* metabolites viewed on the KEGG metabolic network for *T. brucei* (colored pathways) showing putatively identified metabolites (black dots), putative metabolites removed based on predicted retention time (red dots), and putative metabolites removed by noise or MS artifact filters (yellow dots). The retention time filter successfully removed many putative metabolites that were not in the predicted T. brucei metabolome (red dots on gray pathways) while retaining many predicted metabolites (black dots on colored pathways).

3.2.4 Metabolomics example: *T. brucei*

The applicability of predicted retention times to metabolite identification from a real biological metabolomics data set was determined by analysis of cell extracts from *T. brucei*. Filtering by predicted retention time allowed 1382 false identifications (35%) to be removed from the initial output of 3969 putative metabolites from combined PIESI and NIESI data (Supporting Information S7). Application of additional filters to remove noise and MS artifacts resulted in putative identification of 690 metabolites. It is not possible to absolutely confirm the accuracy of these putative identifications in a biological matrix of unknown composition; however, mapping metabolites to organism-specific metabolic pathway reconstructions allows an indication of the usefulness of the results. Excluding lipids and peptides (which are not specifically identified in the KEGG or BioCyc pathway databases), over 60% of the putatively identified metabolites in this study were in agreement with the predicted *T. brucei* metabolome based on KEGG (Kanehisa *et al.*, 2010) and TrypanoCyc (Chukualim *et al.*, 2008) annotations, compared to 39% before filtering (Figure 3.3).

3.2.5 Applications and limitations

While predicted retention times generally do not allow unambiguous metabolite identification, they allow a rapid and efficient mechanism for automated removal of false identifications from metabolomics results, which would otherwise require many hours of interpretation by an experienced analyst. In addition to removal of false identifications, predicted retention times also improve peak annotation when isomeric compounds exist for a specific formula. A total of 78% of metabolites in our database have isomers, highlighting a major inherent limitation of identification by accurate mass, and retention time data assist with disambiguation of these isomers. For example, thymine and imidazole 4-acetate are biologically and structurally distinct isomeric metabolites, which by definition, have identical exact masses but can be easily identified by analysis of retention times (Figure 3.4). It should be noted that isomers often have very similar physicochemical properties and therefore exhibit similar (or even identical) retention times, preventing absolute metabolite identification by retention time prediction in some cases. However, in many cases this approach provides a rapid and efficient method to produce the most likely metabolite annotations based on physicochemical properties, thus increasing the accuracy of biological interpretation of untargeted metabolomics data. Ultimately, absolute identification of specific metabolites requires comparison with authentic standards and additional analytical techniques.

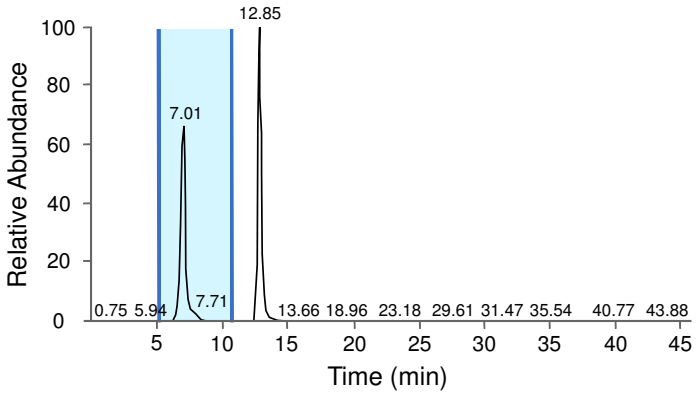


Figure 3.4: Extracted ion chromatogram for $C_5H_6N_2O_2$ (m/z = 126.0429) from solution containing thymine (RT = 7 min) and imidazole 4-acetate (RT = 12.9 min). The predicted retention time window for thymine is shown in blue, demonstrating a role for retention time prediction in isomer identification.

It is expected that retention time drift will prevent direct application of predicted retention times from this model to other platforms, and drift has also been observed on this analytical platform over the lifetime of the column. For this reason, it is suggested that authentic standards are analyzed with each analytical batch and predicted retention times calculated for each batch based on these standards. As it is not practical to routinely analyze every standard independently, we suggest two mixtures for routine analysis, each containing common metabolite standards but avoiding combinations of metabolites that are isomeric with other standards or with MS fragments/adducts of other standards (Supporting Information S1). These authentic standard retention times can be entered into the RTcalculator file in Excel (Supporting Information S2), which includes a macro to recalculate the MLR equation specific to retention times observed in a particular batch and automatically generates predicted retention times for the entire metabolite database.

The QSRR model described herein is specific for analysis with the ZIC-HILIC column and acidic mobile phase (described in Experimental Section). Nevertheless, the general concept of retention time prediction to assist with metabolite identification can be applied to untargeted metabolomics analyses with other chromatographic platforms. To demonstrate this, we analyzed the authentic standards on a ZIC-pHILIC column with alkaline ammonium carbonate mobile phase (Pluskal *et al.*, 2010a). Simply changing the log *D*, positive charge, and negative charge variables to reflect the pH of the mobile phase (pH 9) resulted in a correlation coefficient (r^2) of 0.74 and accurate retention time prediction of 92% of metabolite standards within the 35% RT window. Previous QSRR studies with a variety of compound classes suggest that potential also exists for application of this type of approach to reversed phase LC–MS, for example, the retention-based prediction of log *P* for a diverse range of

compounds (reviewed in Kaliszan (2007)) could be reversed and optimized to allow a log P-based prediction of retention time for hydrophobic metabolites.

3.3 Conclusions

A simple QSRR model for retention time prediction of biological metabolites with ZIC-HILIC chromatography and a user-friendly Excel template allows users to calculate predicted retention times for a database of metabolites based on retention times for authentic metabolite standards. Application of the model leads to markedly improved metabolite annotation with accurate MS-based metabolomics, thus allowing biochemical interpretation of data from untargeted metabolomics studies and bringing metabolomics technology closer to the ultimate goal of global, untargeted analysis of metabolic pathways and systems.

Acknowledgements

D.J.C. gratefully acknowledges financial support from the National Health and Medical Research Council, Australia. Instrumentation was provided by the Scottish Universities Life Sciences Association (SULSA) through the Scottish Metabolomics Facility at the University of Glasgow. A.J. is supported by an NWO-Vidi award to R.B. The authors thank Floriane Laurent for technical assistance.

Supporting Information

Additional information as noted in text. This material is available free of charge via the Internet at `http://pubs.acs.org/doi/suppl/10.1021/ac2021823`.

Chapter 4

Separating the wheat from the chaff: a prioritisation pipeline for the analysis of metabolomics datasets

Andris Jankevics[1, 2], Maria Elena Merlo[1, 3], Marcel de Vries[4], Roel J. Vonk[4], Eriko Takano[3] and Rainer Breitling[1, 2]

Metabolomics 8(Supp. 1):29–36, 2012.

1 Groningen Bioinformatics Centre, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Groningen, The Netherlands
2 Institute of Molecular, Cell and Systems Biology, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, United Kingdom
3 Microbial Physiology, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Nijenborgh 7, 9747 AG Groningen, The Netherlands
4 Centre for Medical Biomics, University Medical Centre Groningen, Groningen, The Netherlands

abstract

Liquid Chromatography Mass Spectrometry (LC–MS) is a powerful and widely applied method for the study of biological systems, biomarker discovery and pharmacological interventions. LC–MS measurements are, however, significantly complicated by several technical challenges, including: (1) ionisation suppression/enhancement, disturbing the correct quantification of analytes, and (2) the detection of large amounts of separate derivative ions, increasing the complexity of the spectra, but not their information content. Here we introduce an experimental and analytical strategy that leads to robust metabolome profiles in the face of these challenges. Our method is based on rigorous filtering of the measured signals based on a series of sample dilutions. Such data sets have the additional characteristic that they allow a more robust assessment of detection signal quality for each metabolite. Using our method, almost 80% of the recorded signals can be discarded as uninformative, while important information is retained. As a consequence, we obtain a broader understanding of the information content of our analyses and a better assessment of the metabolites detected in the analyzed data sets. We illustrate the applicability of this method using standard mixtures, as well as cell extracts from bacterial samples. It is evident that this method can be applied in many types of LC–MS analyses and more specifically in untargeted metabolomics.

Untargeted metabolomics aims to describe living systems by the set of metabolites present in a cell at certain moment of time and under specific environmental constraints (Dettmer *et al.*, 2007; Fiehn, 2002; Oldiges *et al.*, 2007). Since metabolites are the final link between the gene expression and the phenotype exhibited by the cell, metabolomics represents a valuable tool to achieve a better understanding of an organism's phenotype (Fiehn, 2002; Oldiges *et al.*, 2007). The study of the metabolome is complementary to the other "omics" sciences (genomics, transcriptomics, proteomics, fluxomics...) and fits well with the general approach of systems biology (Arita, 2009).

Important advances have been realized in the past years for untargeted metabolite profiling in different research fields, from human health to nutrition (Kamleh *et al.*, 2008; Scalbert *et al.*, 2009). However, metabolomics is still an emerging field in the post-genomic arena. For example, due to the chemical diversity of cellular metabolites and the complexity of the cell extracts, there is no single method which can separate, detect and identify all small molecules present in a cell extract. Fur-

thermore the Achilles' heel of metabolomics remains the identification and structure elucidation of metabolites (Kind and Fiehn, 2010). Sometimes, fragmentation patterns of the molecules can be used for identification. For metabolomics data the detected fragment patterns can, e.g., be matched to online databases, like Metlin (Smith *et al.*, 2005), and assigned to a quality score. But in our experiments we have however observed that the scan time of the LTQ-Orbitrap is considerably affected by the inclusion of fragmentation steps, making the normal LC–MS data stream fragmentary and difficult to analyze automatically. As more convenient alternative, the Orbitrap Exactive platform (without the linear iontrap but with faster scan speeds) can be used to capture more data points using the positive–negative polarity switch mode (Lu *et al.*, 2010). Thus, currently matching on mass alone to databases is the most commonly used method. Unfortunately, this approach to metabolite identification is very seriously hampered by the fact that the vast majority of the signals in the data set can be caused by contaminants in the sample or LC–MS system (Keller *et al.*, 2008), technical artefacts and so-called "derivative peaks" (Scheltema *et al.*, 2009). In many cases, several peaks or signals share the same identifications, even if signals are detected with an accuracy of better than 2 ppm, as is routinely possible using, e.g., modern Fourier Transform mass spectrometers, like the Orbitrap (Scheltema *et al.*, 2008). Such spurious peaks need to be checked manually and assigned to their real identification or discarded if the signal shows typical artefacts.

Our goal was to develop an analytical method that would be able to eliminate a substantial part of the spurious signals from the data set. This required the development of new approaches and the collection of an unusual type of data on biological samples and mixtures of analytical standards, to distinguish real effects from spurious fluctuations in LC–MS analyses and peak detection algorithms. The strategies developed here will be generally useful for metabolomics.

4.1 Materials and methods

4.1.1 Amino acid standard mixture samples

A mixture of 38 physiological amino acid standards (Product No. A9906, Sigma) was used. In the stock solution, amino acids and related compounds are contained at a final concentration of 0.5 μmol/ml $\pm$ 4% in 0.2 N lithium citrate buffer, pH 2.20, containing thiodiglycol (2% w/v) and phenol (0.1% w/v) as antioxidant and preservative, respectively. The concentration in the injected diluted samples is described in Table 4.1.

Dilution factor	1/8	1/16	1/32	1/64	1/128	1/256	1/512	1/1024
Concentration (μmol/ml)	0.0625	0.0313	0.0156	0.0078	0.0039	0.0020	0.0010	0.0005
Injected on column (pmol)	312.5000	156.2500	78.1250	39.0625	19.5313	9.7656	4.8828	2.4414

Table 4.1: Dilution factor and concentrations of the analysed samples.

4.1.2 Biological samples

Analytical samples were obtained from *Streptomyces coelicolor* wild-type M145 strain (Bentley *et al.*, 2002). Bacteria were grown in 50 ml liquid minimum medium (Nieselt *et al.*, 2010) as described (Takano *et al.*, 2001). Cells from 25 ml of culture were collected on a 0.45 μm filter by vacuum filtration and washed twice with 25 ml of 2.63% NaCl solution. For cell quenching, the filter with the collected cells was quickly moved into 60% methanol solution (HPLC-grade, Boom, The Netherlands) pre-chilled at –20°C and frozen in liquid nitrogen. Samples were stored at –80°C until metabolite extraction was performed.

Metabolites were extracted by three freeze-thaw cycles. Cells were thawed in an ethanol bath at –20°C (15 min), vortexed vigorously for 1 min and, right afterwards, frozen in liquid nitrogen for 5 min. The cycle was repeated three times. After the third cycle, the samples were centrifuged at 4500 rpm for 10 min at –9°C. The supernatant (cell extract) was collected and stored at –80°C until LC–MS analysis. Before analysis, obtained samples were diluted with the same dilution factor as for the analytical standards mixture, resulting in eight samples with different metabolite concentrations.

4.1.3 LC–Orbitrap MS analysis

The analytical mixtures and cell extracts were analyzed by liquid chromatography coupled to a high-accuracy LTQ Orbitrap XL mass spectrometer (Thermo Fisher Scientific, Germany).

Two chromatographic columns were used: a reversed-phase Shim-pack XR-ODS C18 column (Achrom, Belgium) (3.0 × 75 mm, 2.2 μm, Shimadzu Corp.) and a ZIC-HILIC column (Achrom, Belgium) (150 × 2.1 mm, 3.5 μm, Merck Sequant AB) fitted with a ZIC-HILIC PEEK guard column (Achrom, Belgium) (15 × 1.0 mm; 5 μm, Merck Sequant AB).

For the C18 column, the flow rate was set to 0.6 ml/min; the mobile phase consisted of (A) 0.1% formic acid in water and (B) 0.1% formic acid in acetonitrile.

A gradient of 18 min was used. The elution of solvent B started at 2% for the first 2 min and was increased to 95% within 8 min. This composition was maintained for 2 min, after which the elution of B was decreased to 2% within 1 min. To re-equilibrate the system, the elution of B was held at 2% for 5 min.

For the ZIC-HILIC column, the flow rate was set to 0.1 ml/min; as buffers, (A) 0.1% formic acid in acetonitrile and (B) 0.1% formic acid in water were used. A gradient of 40 min was applied. Solvent A was set to 80% as starting condition. The elution fraction of solvent B was increased to 40% within 6 min and maintained at 40% for 12 min, after which solvent B was increased to 90% in a 4 min-interval. This composition was held for 2 min after which B was decreased to 20% in 2.5 min. The gradient was held at 20% B for 13.5 min to re-equilibrate the system.

The sample volume injected was 5 μl for both columns, and two technical replicates were recorded for the C18 analysis, and three replicates on the HILIC column.

The system was operated with the electrospray ionization source in positive mode. Full-scan spectra were obtained over an m/z range of 50–1000 Da.

ULC grade acetonitrile, formic acid and water were purchased at Biosolve (Netherlands).

4.1.4 Data processing

Raw data files from the mass spectrometer were converted into the mzXML format by the ReAdW.exe utility (a tool of the Trans-Proteomic Pipeline software collection, downloaded from `http://tools.proteomecenter.org/wiki/index.php?title=Software:ReAdW`).

The CentWave (Tautenhahn *et al.*, 2008) feature detection algorithm from the XCMS (Smith *et al.*, 2006) package was used on each individual data file. Further processing was handled by the flexible data processing pipeline mzMatch (Scheltema *et al.*, 2011), performing noise removal (Windig, 2004) and several steps of signal filtering and peak matching. The first matching step involved aligning of the chromatographic features between technical replicates of a single sample. Peaks that were not detected in all technical replicates were discarded from further analysis. In the second matching step, the chromatographic peaks, which were combined in single files containing technical replicates in the previous matching step, were aligned to each other for all eight dilutions. After combining the eight measurements in a single file, there were still peak sets that did not include peaks from every sample. Such gaps were filled by extracting ion chromatograms within the retention time and mass window of the given peak set directly from the raw data files.

Derivative signals (isotopes, adducts, dimers and fragments) were automatically

annotated by correlation analysis on both signal shape and intensity pattern, as described (Scheltema *et al.*, 2009). These peaks were not discarded and their assigned annotations were taken into account in the subsequent analysis.

Putative identifications were made by matching the detected masses to a database of Streptomyces coelicolor (ScoCyc) metabolites, a contaminants database (Keller *et al.*, 2008), and the list of analytical standards in the standard mixture. The metabolite database was obtained from a genome annotation file created by Jonathan Moore as part of the SysMO STREAM project (`https://www.wsbc.warwick.ac.uk/groups/sysmopublic/`), which is also available for download from the BioCyc project page (Karp *et al.*, 2009) as a flat-file in Pathway Tools format (Karp *et al.*, 2002).

Pearson's correlation of binary logarithm of the peak intensities was applied to evaluate dilution trends in the obtained data set. Samples for the 8 dilution points were ordered from highest to lowest concentration, so that metabolites matching the sample dilution trend would show high negative correlation values between intensity and sample number. Correlation values smaller than –0.85 were considered as indicating a significantly reproducible dilution trend.

For low-abundance peaks, where signals for the highest dilutions were below the limit of detection, correlation values were calculated for the detectable consecutive measurements (at least 3 dilution points were required).

All statistical analyses and graphical routines were handled in R R Core Team (2012).

Raw data files in mzXML format, R code containing the complete data processing pipeline, as well final peak tables are available for download at `http://mzmatch.sourceforge.net/metabolomics.html`.

4.2 Results and discussion

Our study was carried out in two steps. First we wanted to validate our filtering method by applying it to the data sets of the mixtures of analytical standards. The resulting numbers of detected peaks are shown in Table 4.2. Data for both chromatographic columns are shown: even for relatively simple samples (39 compounds in the mix of standards) a huge amount of the peaks were detected (2831 peak sets for C18 data, and 11169 for HILIC). Only about 20–30% of these signals can be identified in chemical databases or assigned to known contaminants.

A significant amount of the uninformative signals could be removed after application of the dilution trend filter. For example, in the unfiltered data set for HILIC

	C18				HILIC			
	BF		Filtered		BF		Filtered	
	1	2	1	2	1	2	1	2
Detected as standards (bp)	49	12	11	10	409	28	91	26
Detected as standards (rp)	30	27	20	20	256	30	99	28
Detected as contaminants (bp)	69	23	1	1	227	28	13	8
Detected as contaminants (rp)	40	17	5	4	147	22	22	9
Detected in ScoCyc (bp)	94	17	6	2	516	68	70	29
Detected in ScoCyc (rp)	72	37	24	23	383	92	115	58
Unidentified (bp)	1335		106		4745		493	
Unidentified (rp)	1142		348		4486		1337	

Table 4.2: Comparison of number of the peaks extracted for the standard mixtures samples. The fraction of features uniquely identified as standard -compounds is significantly increased after application of trend filtering. BF – Before filtering. bp – labelled as base peaks by mzMatch software, rp – labelled as derivative peaks, 1 – number of peaks, 2 – number of unique identifiers.

data 28 unique standard compounds were matching 409 features within 5 ppm mass accuracy window. After application of the dilution trend filter, this number decreased to 91 features matching 26 unique standard compounds. In other words, the number of detected compounds is not significantly changing, while the number of total peaks in the data set is decreasing by almost 5 times and the number of unambiguous matches is substantially increased (Fig. 4.1a). Manual inspection showed that the two putative standard compounds removed by application of the filter were artefacts, i.e. these two compounds were not really detectable. Also, a very large amount of the signals matching the ScoCyc database (which should not be present in samples of analytical standards) was removed by the trend filter, as were most of the unidentifiable compounds, which also do not match the expected composition of the samples. Overall the fraction of correctly identifiable compounds is dramatically increased.

A list of the standard compounds detected on both C18 and HILIC columns is shown in Table 4.3. The following structural isomers could not be distinguished: L-alanine, L-sarcosine and β-alanine; γ-amino-N-butyric acid, D,L-β-aminoisobutryic acid and L-α-amino-N-butyric acid. For L-isoleucine/L-leucine and 1-methyl-L-histidine/3-methyl-L-histidine two peaks eluting close to each other were observed. Ammonium chloride was not detected on either column (because of its low molecular weight), and L-ornithine was not detected on the HILIC column. Almost no separation was achieved on the C18 column (most of the signals eluted within the first minute of the analytical run). Surprisingly high quantification accuracy (correlation value is close to –1, i.e. a linear relationship between intensity and sample dilution) can be observed for

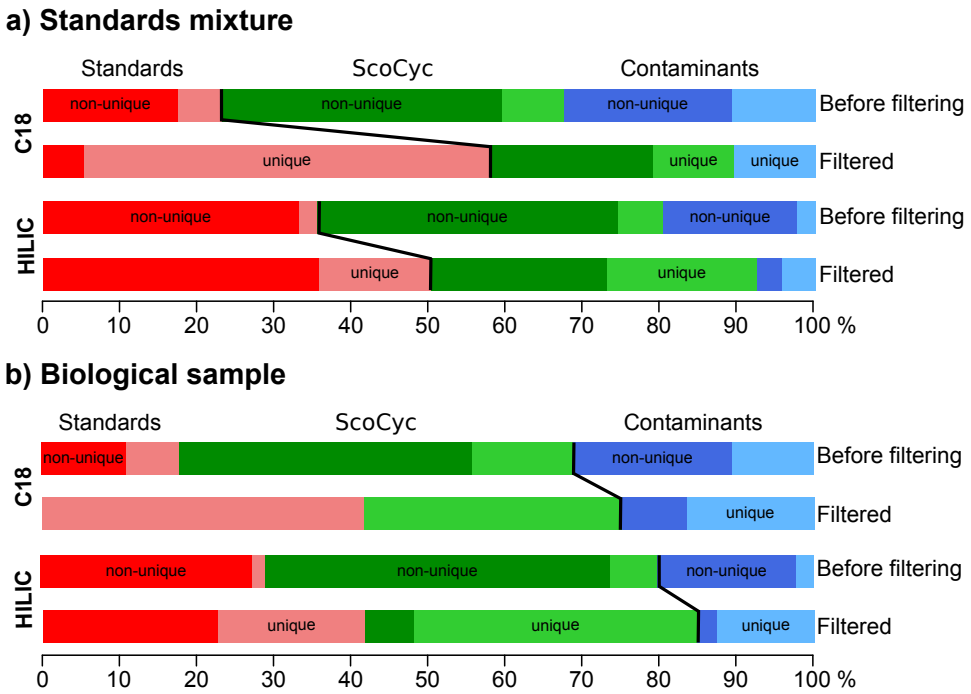


Figure 4.1: Proportional relationship between identified compounds before and after filtering on dilution trend. Compounds labelled as base peaks by the mzMatch software are shown. For the standards mixture (**a**) where only matches to the standard compounds are expected, a clear increase of the fraction of identified peaks can be seen after filtering. Importantly, the fraction of uniquely identified compounds (lighter shade of the color) is also strongly increasing. In other words, after filtering more compounds with unambiguous, unique identifications are retained. The same trend can be also seen in the data for the biological samples (**b**), where matches to the standard compounds and the ScoCyc data base are expected. Matches to the contaminant compounds decrease in the filtered data, and the number of unique identifications increases substantially.

almost all analytical standards on both chromatographic columns.

The resulting numbers of detected peaks after processing of biological samples are shown in Table 4.4. Surprisingly, the amount of detected peaks is comparable to the numbers seen for the analytical standards, both in the filtered and unfiltered data sets. For the HILIC data set, 639 features were putatively identified in the ScoCyc database (78 unique compounds), but only 28 peaks (24 unique identifiers) were retained after application of the dilution trend filtering. Clear trends in improvement of the data set quality are shown in Fig. 4.1b.

Metabolite (KEGG ID)	Molecular formula	Monois. mass	C18		HILIC	
			Corr.	RT	Corr.	RT
Urea (C00086)	CH_4N_2O	60.03240	–0.88	0min 40s	–0.99	8min 33s
Ethanolamine (C00189)	C_2H_7NO	61.05280	–0.94	0min 32s	–0.99	20min 56s
Glycine (C00037)	$C_2H_5NO_2$	75.03200	–0.89	0min 35s	–1	17min 58s
L-Alanine (C00041)	$C_3H_7NO_2$	89.04770	–0.97	0min 35s	–0.98	14min 58s
γ-Amino-*N*-butyric acid (C00334)	$C_4H_9NO_2$	103.06330	–0.91	0min 37s	–0.90	13min 56s
L-Serine (C00065)	$C_3H_7NO_3$	105.04260	–0.88	0min 35s	–1	18min 05s
L-Creatinine (C00791)	$C_4H_7N_3O$	113.05890	–0.97	0min 34s	–0.85	14min 44s
L-Proline (C00148)	$C_5H_9NO_2$	115.06330	–0.92	0min 38s	–0.98	14min 26s
L-Valine (C00183)	$C_5H_{11}NO_2$	117.07900	–0.99	0min 48s	–0.92	13min 18s
L-Threonine (C00188)	$C_4H_9NO_3$	119.05820	–0.95	0min 35s	–0.89	18min 16s
Taurine (C00245)	$C_2H_7NO_3S$	125.01470	–0.86	0min 36s	–0.99	15min 01s
Hydroxy-L-proline (C01157)	$C_5H_9NO_3$	131.05820	–0.96	0min 36s	–0.85	15min 23s
L-Isoleucine (C00407)	$C_6H_{13}NO_2$	131.09460	–0.99	1min 34s	–1	11min 48s
L-Ornithine (C00077)	$C_5H_{12}N_2O_2$	132.08990	–0.95	0min 28s		
L-Aspartic acid (C00049)	$C_4H_7NO_4$	133.03750	–0.90	0min 36s	–1	16min 39s
L-Lysine (C00047)	$C_6H_{14}N_2O_2$	146.10550	–0.95	0min 28s	–1	30min 7s
L-Glutamic acid (C00025)	$C_5H_9NO_4$	147.05320	–0.92	0min 36s	–0.90	15min 41s
L-Methionine (C00073)	$C_5H_{11}NO_2S$	149.05100	–0.99	1min 02s	–1	12min 48s
L-Histidine (C00135)	$C_6H_9N_3O_2$	155.06950	–0.89	0min 29s	–1	29min 19s
δ-Hydroxylysine (C01211)	$C_6H_{14}N_2O_3$	162.18700	–0.95	0min 28s	–1	30min 18s
L-Phenylalanine (C00079)	$C_9H_{11}NO_2$	165.07900	–0.99	3min 42s	–1	11min 16s
1-Methyl-L-histidine (C01152)	$C_7H_{11}N_3O_2$	169.08510	–0.91	0min 31s	–1	29min 38s
L-Arginine (C00062)	$C_6H_{14}N_4O_2$	174.11170	–0.95	0min 32s	–1	30min 10s
L-Citrulline (C00327)	$C_6H_{13}N_3O_3$	175.09570	–0.93	0min 36s	–1	18min 35s
L-Tyrosine (C00082)	$C_9H_{11}NO_3$	181.07390	–0.99	1min 40s	–1	13min 42s
L-Tryptophan (C00078)	$C_{11}H_{12}N_2O_2$	204.08990	–0.99	4min 39s	–0.99	12min 01s
L-Cystathionine (C02291)	$C_7H_{14}N_2O_4S$	222.06740	–0.87	0min 35s	–1	26min 00s
L-Carnosine (C00386)	$C_9H_{14}N_4O_3$	226.10660	–0.88	0min 28s	–1	31min 01s
L-Cystine (C00491)	$C_6H_{12}N_2O_4S_2$	240.02380	–0.88	0min 35s	–1	25min 17s
L-Anserine (C01262)	$C_{10}H_{16}N_4O_3$	240.12220	–0.98	0min 30s	–1	30min 53s
L-Homocystine (C01817)	$C_8H_{16}N_2O_4S_2$	268.05510	–1	0min 44s	–1	24min 02s

Table 4.3: Identified compounds in the analytical mixture. *Monois. mass* – Monoisotopic mass, *Corr.* – Pearson's correlation coefficient between sample number and the logarithm of the signal intensity, *RT* – retention time

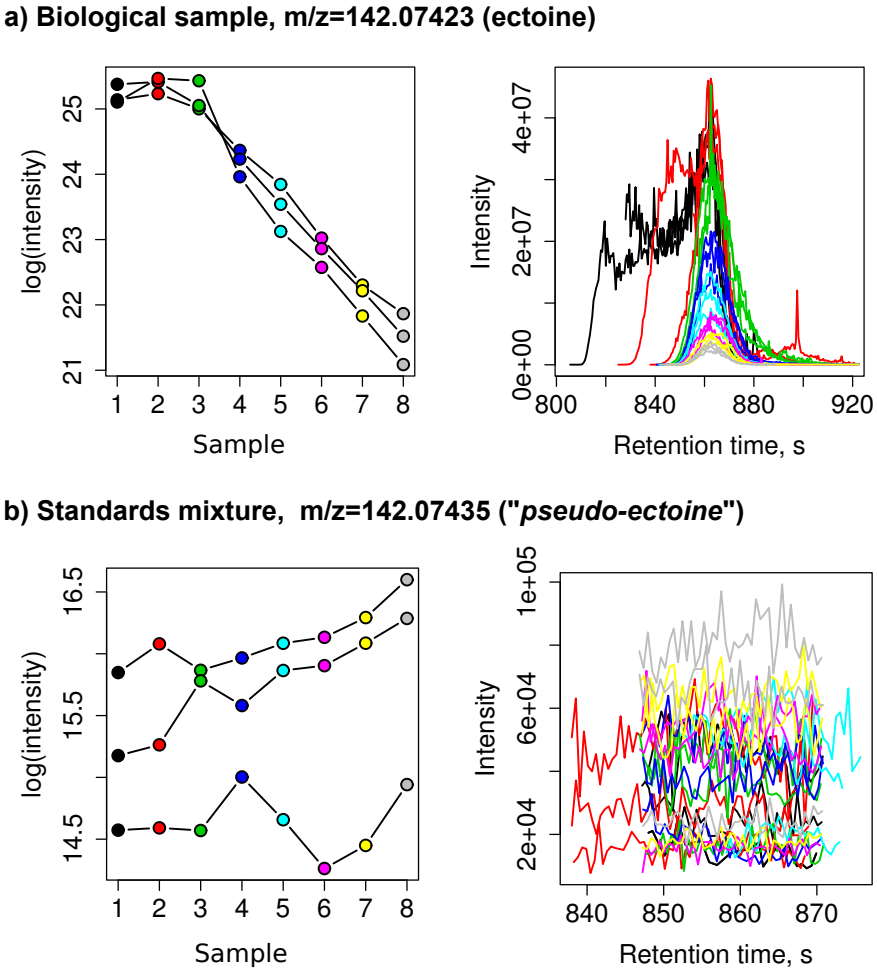


Figure 4.2: Example of the dilution trends (on the left) and extracted mass chromatograms (on the right) for a metabolite putatively identified as ectoine. For the biological samples, which are expected to contain ectoine (Kol *et al.*, 2010), three technical replicates show clearly identifiable dilution trend (trend correlation value –0.97). For the standard mixture, which does not contain ectoine, a random trend is seen in all replicates for the signal putatively identified as ectoine (mass error 0.86 ppm); this putative technical artefact can thus be removed by the trend filtering.

Interesting compounds that were identified (and expected) only in the biological samples on both chromatographic columns are the osmoregulator compound ectoine and hypoxanthine. In Fig. 4.2, an example of dilution trends and chromatographic

peaks for the biological sample (Fig. 4.2a) and the standard mixture (Fig. 4.2b) is given. In both data sets, a peak was identified as matching the mass of ectoine with an apparent mass error less than 1 ppm, but in the standard mixture (which does not contain ectoine), this peak was successfully discarded by the trend filter, as the signal intensity patterns (shown in the left panel of the plot) are not following the sample dilution trend.

	C18				**HILIC**			
	BF		**Filtered**		**BF**		**Filtered**	
	1	**2**	**1**	**2**	**1**	**2**	**1**	**2**
Detected as standards (bp)	34	13	5	5	366	22	32	15
Detected as standards (rp)	16	12	8	8	208	23	20	10
Detected as contaminants (bp)	59	20	3	2	254	28	10	8
Detected as contaminants (rp)	29	16	4	4	129	29	16	12
Detected in ScoCyc (bp)	97	25	4	4	639	78	28	24
Detected in ScoCyc (rp)	36	22	7	7	362	78	46	33
Unidentified (bp)	1235		19		4962		146	
Unidentified (rp)	632		123		3053		359	

Table 4.4: Comparison of the number of the peaks extracted for the biological samples before and after trend filtering. The fraction of compounds with putative identifications is significantly increased after application of the trend filter. BF – Before filtering. bp – labelled as base peaks by mzMatch software, rp – labelled as derivative peaks, 1 – number of peaks, 2 – number of unique identifiers.

The biological samples used in this illustrative example are particularly challenging, due to a large number of peaks with low signal intensities. Our results show that even for such difficult data, the dilution trend filter can be applied with no real danger of losing information of interest. It is also quite obvious that sample dilution factors should be adjusted according to the expected overall metabolite levels in the analysed samples, to avoid over-dilution and loss of signals of interest. To avoid the problem of large correlations occurring by chance when the number of observations is low, the statistical significance of the observed correlation can be examined and the obtained p-values can be used to determine the threshold for peak selection. This method can also be integrated with a quality control sample approach (Sangster *et al.*, 2006), where repeated injection of a pooled randomized sample throughout the analysis serves as a reference for quality control; this approach is commonly used in large populations studies (Zelena *et al.*, 2009). This control sample can be replaced with injections of pooled dilution samples in randomized order. Thereby, without increasing the number of injections for a typical analytical sample batch, it will be possible to simultaneously assess machine stability (as the dilution trend should stay constant) and do a filtering of the data set on highly reproducible signals.

The method suggested here is therefore a useful complement to the commonly used relative standard deviation (RSD) filters (Scheltema *et al.*, 2008; Shah *et al.*, 2000) and the CoDA-DW filters, (Windig, 2004), allowing automatic retrieval of signals of interest, reducing the complexity of the data and consequently speeding up the interpretation process.

The dilution filtering approach can be easily integrated in a complete data processing pipeline (based on mzMatch and XCMS software tools) and used in a semi-automated manner. This is illustrated in the R script provided as supplementary material for this study (`http://mzmatch.sourceforge.net/metabolomics.html`).

4.3 Concluding remarks

We have been able to demonstrate the effectiveness and reliability of a relatively simple data filtering strategy. The proposed trend correlation filter significantly decreases the amount of non-informative signals in the data sets and makes metabolite identification much easier. We could show that even very stringent filtering of the data is not causing a loss of informative signals.

Our illustrative application to biological samples demonstrates that our approach can also be applied to assess the performance of metabolite extraction from the samples. This allows a more reliable estimate of the true metabolomic complexity observed in a particular experiment.

Acknowledgements

The authors gratefully acknowledge the contributions of Richard Scheltema (Max Planck Institute for Biochemistry, Germany), Ruben t'Kindt (Metablys, Belgium) and Darren Creek (University of Glasgow, UK) during many discussions on data processing and mass spectroscopy-related topics. The authors have declared that no competing interests exist. AJ is supported by an NWO-Vidi award to RB. MEM is funded by a 4 × 4 Ubbo Emmius scholarship and ET by a Rosalind Franklin Fellowship, both from the University of Groningen. RJV was supported by an investment grant from NWO.

Chapter 5

mzMatch–ISO: an R tool for the annotation and relative quantification of isotope-labelled mass spectrometry data

Achuthanunni Chokkathukalam[1*], Andris Jankevics[1,2*], Darren J. Creek[3,4], Fiona Achcar[1], Michael P. Barrett[4], Rainer Breitling[1,2,5]

Bioinformatics 29(2):281–283, 2013.

1 Institute of Molecular Cell and Systems Biology, College of Medical Veterinary and Life Sciences, University of Glasgow, Glasgow, UK.
2 Groningen Bioinformatics Center, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Groningen, The Netherlands
3 Department of Biochemistry and Molecular Biology, Bio21 Molecular Science and Biotechnology Institute, University of Melbourne, Melbourne, Australia
4 Wellcome Trust Centre for Molecular Parasitology, Institute of Infection, Immunity and Inflammation, College of Medical Veterinary and Life Sciences, University of Glasgow, Glasgow, United Kingdom
5 Faculty of Life Sciences, Manchester Institute of Biotechnology, University of Manchester, Manchester, United Kingdom
* Equal contribution

abstract

Motivation: *Stable isotope labelling experiments have recently gained increasing popularity in metabolomics studies, providing unique insights into the dynamics of metabolic fluxes, beyond the steady–state information gathered by routine mass spectrometry. However, most liquid chromatography–mass spectrometry data analysis software lacks features that enable automated annotation and relative quantification of labelled metabolite peaks. Here we describe mzMatch–ISO, a new extension to the metabolomics analysis pipeline mzMatch.R.*

Results: *Targeted and untargeted isotope profiling using mzMatch–ISO provides a convenient visual summary of the quality and quantity of labelling for every metabolite through four types of diagnostic plots that show a) the chromatograms of the isotope peaks of each compound in each sample group; b) the ratio of mono–isotopic and labelled peaks indicating the fraction of labelling; c) the average peak area of mono–isotopic and labelled peaks in each sample group; and d) the trend in the relative amount of labelling in a predetermined isotopomer. To aid further statistical analyses, the values used for generating these plots are also provided as a tab–delimited file. We demonstrate the power and versatility of mzMatch–ISO by analysing a ^{13}C labelled metabolome dataset from trypanosomal parasites.*

Availability: *mzMatch.R and mzMatch–ISO are available free of charge from* `http://mzmatch.sourceforge.net` *and can be used on Linux and Windows platforms running the latest version of R.*

Liquid chromatography–mass spectrometry (LC–MS) is a technique that combines the physical separation capabilities of liquid chromatography with the highly sensitive mass detection properties of mass spectrometry. Metabolomics studies utilise LC–MS for the global detection and relative quantification of metabolites in complex biological samples. Recently, LC–MS has been applied to trace the metabolism of stable isotope labelled metabolic precursors in biological systems as a function of time (Supplementary Figure S1) (Chaneton *et al.*, 2012; Le *et al.*, 2012). Such experiments can provide unique insights into the dynamics of metabolic fluxes, beyond the steady–state information gathered by routine metabolomics.

Stable isotope labelled metabolites possess the same chromatographic properties as their unlabelled counterparts and can be readily identified from the MS dataset based on their expected mass (Figure 5.1). However, a key challenge that metabolomics researchers face is the limited number of suitable bioinformatic solutions for metabolome–wide isotope labelled data analyses. Multiple MS data analysis tools

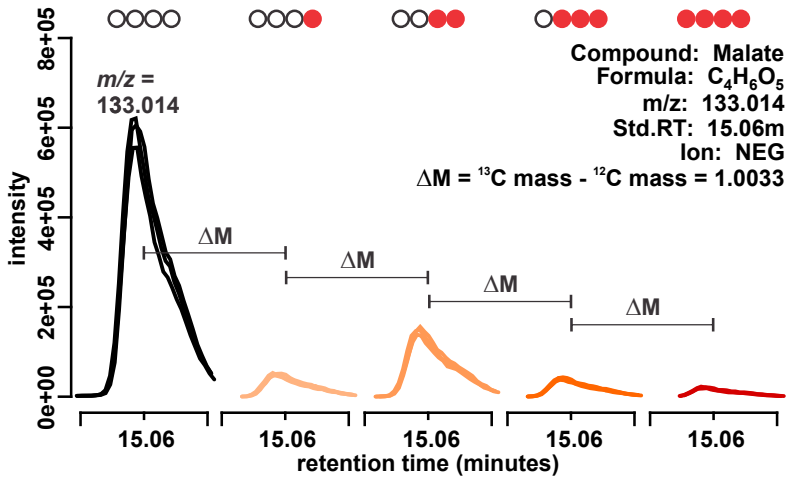


Figure 5.1: A representative example of an unlabelled peak (black) and its corresponding ^{13}C labelled peaks. These peaks elute at the same retention time, but their masses differ by the difference in the mass of heavy and light carbon. Red circles show the number of labelled carbons that each peak represent.

are available (Dunn *et al.*, 2012), including widely used open source software such as mzMine (mzmine.sourceforge.net), mzMatch (mzmatch.sourceforge.net) and XCMS (metlin.scripps.edu), and commercial software such as SIEVE (www.thermo.com), MassHunter (www.chem.agilent.com) and MarkerLynx (www.waters.com). While they are all capable of identifying and quantifying metabolites of interest in unlabelled data, features that enable the extraction and relative quantification of isotope peaks from labelled data either require manual intervention or are non–existent. Furthermore, software that can handle labelled MS data, such as MetExtract (Bueschl *et al.*, 2012) and MAVEN (Melamud *et al.*, 2010), lacks appropriate peak picking algorithms in the processing pipeline. The ability to generate a comprehensive graphical visualisation of the output along with the extensible and scriptable nature of the software itself makes mzMatch–ISO a unique data analysis tool for isotope labelling studies.

Here we present a novel and unique tool, mzMatch–ISO, that circumvents the bottlenecks described above by performing fully automated targeted or untargeted annotation and relative quantification of mono–isotopic and corresponding isotope labelled peaks of metabolites in stable isotope labelled LC–MS data to generate plots and tables that describe the labelling pattern in detail. mzMatch–ISO is an extension to mzMatch, an open–source Java toolbox for MS data processing and visu-

alization (Scheltema *et al.*, 2011). Features of mzMatch—enabled by the R package mzMatch.R—including the new PeakML data exchange format and the data analysis pipeline were described by Scheltema *et al.* (2011). mzMatch has been applied to many metabolomics data analyses (Jankevics *et al.*, 2012; t'Kindt *et al.*, 2010a), and is the underlying platform for software such as IDEOM (Creek *et al.*, 2012a). Currently, only LC–MS data analysis is supported by mzMatch–ISO; however, it is possible to process gas chromatography–mass spectrometry data analyses using mzmatch.R and mzmatch–ISO with some additional scripting.

5.1 Methods

Isotope profiling using mzMatch–ISO requires the LC–MS raw data (.mzXML) files to be pre–processed by extracting the peaks using XCMS (Smith *et al.*, 2006) and aligning, noise filtering and gap–filling peaks into a combined PeakML file containing all samples using mzMatch.R. In addition to the PeakML file, targeted profiling requires an additional tab–delimited input text file containing the list of compounds of interest (see Supplementary Figure S2). This list can also contain characteristic adducts or fragments of metabolites of interest where appropriate. Automated targeted and untargeted isotope profiling can be performed using the mzMatch–ISO functions PeakML.Isotope.TargettedIsotopes() and PeakML.Isotope.UntargettedIsotopes(), respectively. The latter can be used for profiling global label distribution by looking for the isotopes of all identified peaks in a PeakML file; or all compounds in databases such as KEGG (Kanehisa *et al.*, 2012) or HMDB (Wishart *et al.*, 2009); or using common metabolic transformations recursively, as described in Breitling *et al.* (2006a), Gipson *et al.* (2008), Rogers *et al.* (2009) and Weber and Viant (2010). All parameters used in these functions are described at `http://mzmatch.sourceforge.net/isotopes-targetted.php`.

For both targeted and untargeted analysis, mzMatch–ISO generates two outputs—a PDF file and a tab–delimited file. The former contains one page per metabolite with various plots that describe the pattern of labelling observed (Figure 5.2). The page header shows compound information from the target list or the database used for identification, and the ionisation polarity (Figure 5.2a). In cases where more than one peakset is present within a given mass window, each peakset is plotted on a separate page of the PDF file (Figure 5.2b); usually the correct peakset can be identified by considering the retention time and intensity profile.

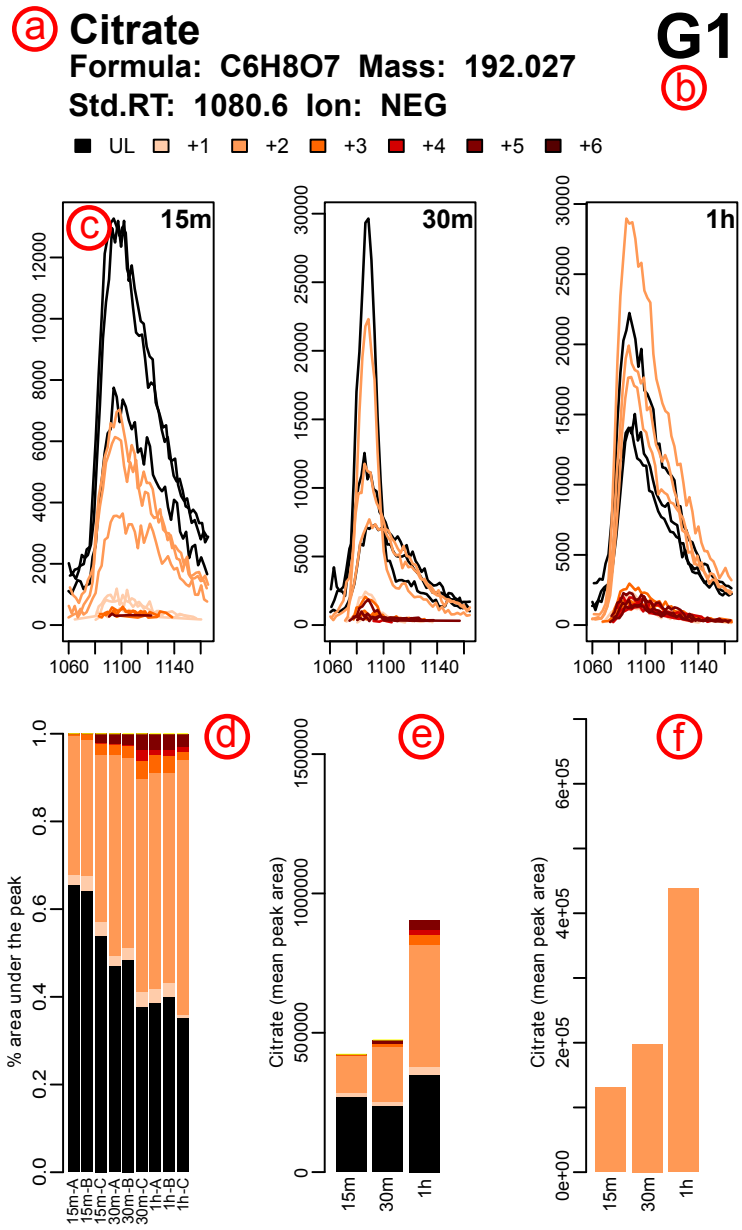


Figure 5.2: Representative example of the output PDF file generated by mzMatch–ISO (see text for details).

Chromatograms of each unlabelled peak and its corresponding labelled isotopomers for each sample in each condition are shown (Figure 5.2c), the peak area/height being stored in the accompanying tab–delimited file. The intensity and shape of the chromatograms helps to assess the effect of noisy or incomplete peaks on the reported pattern of labelling. Furthermore, these chromatograms can be used along with the plot in Figure 5.2d, showing the normalized peak area/height of each mono–isotopic peak and its isotopomers in each replicate, to make informed decisions on outliers by assessing the variability in labelling between replicates. The overall trend in the labelling pattern of a metabolite, as observed between various conditions involved in the study, is also visualised (Figure 5.2e). This plot is especially useful in time–series analyses to rapidly observe the dynamics of relative concentration changes. The final plot (Figure 5.2f) not only highlights the labelling trend of an isotopomer of choice, but, in the case of one–carbon labelling studies it also compares observed signals with the theoretical intensity expected based on the natural abundance of the relevant isotope. This plot is of extreme importance in studies where natural abundance has to be distinguished from low levels of labelling.

5.2 Results

The automated untargeted isotope annotation and relative quantification capabilities of mzMatch–ISO are demonstrated by the analysis of LC–MS data from procylic form *Trypanosoma brucei* grown on ∼50% uniformly ^{13}C–labelled glucose medium for five days. The plot in Supplementary Figure S3 generated from the tab–delimited output file (data are provided in Supplementary file S4 and the scripts are available on the website) highlights the capabilities of mzMatch–ISO in demonstrating a complex biological phenomena.

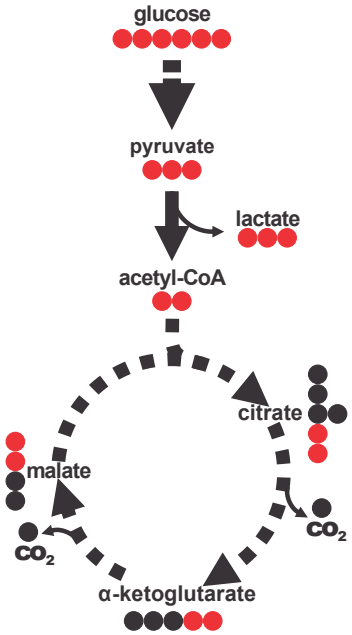
mzMatch–ISO provides an efficient and user–friendly output for the analysis and compact visualization of isotope labelled metabolomics datasets without the need for specialist bioinformatics skills, allowing rapid, precise and meaningful biological interpretation. The algorithm can be implemented directly in R, or from the IDEOM graphical user interface, to facilitate follow–up statistical processing, analyses and re–plotting of the results.

Acknowledgements

AC was funded by a Scottish Universities Life Sciences Alliance (SULSA) grant to RB. Funding for AJ was provided by a Netherlands Organisation for Scientific Research NWO-VIDI grant to RB. DJC was supported by an Australian NHMRC postdoctoral training fellowship. FA was supported by SysMO, NWO-Vidi and SULSA. MPB was supported by the Wellcome Trust through The Wellcome Trust Centre for Molecular Parasitology, which is supported by core funding from the Wellcome Trust [085349].

Supporting Information

Additional information as noted in text. "Supplementary file 4" is available free of charge via the Internet at `http://bioinformatics.oxfordjournals.org/content/suppl/2012/11/19/bts674.DC1/Supplementary_S4.xls`.

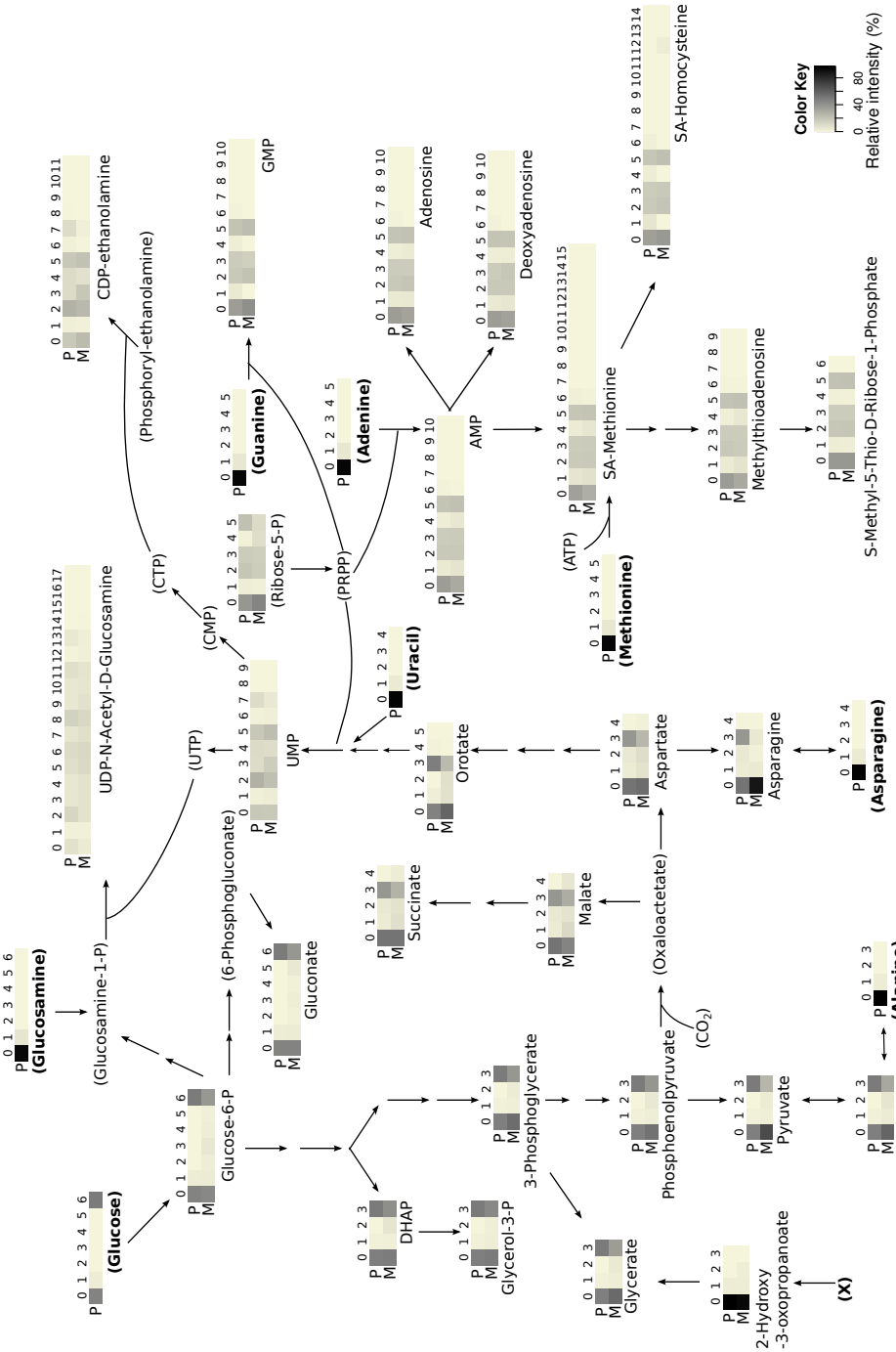


Supplementary Figure S1: Tracing the distribution of heavy isotope labelling in metabolites constituting a pathway. Red and black filled circles represent labelled ^{13}C carbon and unlabelled ^{12}C carbon, respectively.

id	name	formula	rt	follow
GLU	Glucose	C6H12O6	13.56	6
G6P	Glucose-6-phosphate	C6H13O9P		6
GAP	Glyceraldehyde-3-phosphate	C3H7O6P		3
⋮	⋮	⋮		

Supplementary Figure S2: Tab–delimited input file required for targeted isotope profiling containing the details of the compounds of interest. A unique identifier that distinguishes the compound, the name of the compound and its formulae as shown in the first three columns are mandatory. Providing standard retention time (in the 'rt' column), if available, can aid in the precise identification of mono–isotopic peaks and corresponding isotopic peaks for the compound of interest. To enable this feature an appropriate retention time window has to be specified in the 'stdRTWindow' parameter of the PeakML.Isotope.TargettedIsotope() function. The last column, 'follow', can be used to specify the isotopomer that has to be tracked by generating a separate plot in the PDF output (6: track the isotopomer with 6 heavy atoms, etc.)

Supplementary Figure S3 (figure shown on the next page): A flux map generated as a result of an automated untargeted annotation and quantification of isotope profiles of compounds in an LC–MS dataset from procylic form *Trypanosoma brucei* grown on ∼50% uniformly ^{13}C-labelled glucose medium for five days. The good match between expected (P) and observed (M) isotopomer patterns confirm the performance achieved by the automated mzMatch–ISO pipeline. Raw data from LC–MS experiments were initially converted to .mzXML files using the ReAdW tool from the Trans-Proteomic Pipeline (TPP) and then subjected to the automated mzMatch processing pipeline as described in `http://mzmatch.sourceforge.net/tutorial.mzmatch.r.advanced.php`. The PeakML file after combining based on conditions, filtering and gap–filling was profiled for isotopes of all compounds in the KEGG database using the PeakML.Isotope.UntargettedIsotopes() function using a 5–ppm mass window. Data from the tab–delimited output file was used to generate the heat map for the measured values in the plot. The predicted labelling patterns of the metabolites were calculated based on a binomial distribution considering the expected maximal number of atoms obtained from glucose and from other unlabelled sources for each metabolite. The glucose pool was considered to contain 50% labelled and 50% unlabelled glucose (with natural abundance of ^{13}C) (Murphey and Nier, 1941). Ribose 5-phosphate (R5P) is obtained from the pentose phosphate pathway (PPP) and hydrolysis of unlabelled nucleosides present in the medium. The complex exchanges of carbons in the PPP make it difficult to predict the R5P labelling pattern. Therefore, the predicted pattern of R5P was based on the measured pattern of S-Methyl-5-Thio-D-Ribose 1-Phosphate, the labelled carbons of which, can only come from R5P. Subsequently, the labelling patterns of all metabolites derived from R5P were calculated based on this pattern. Metabolites in bold letters are external, i.e., they do not contain atoms derived from labelled glucose, but only the natural abundance of ^{13}C.



Supplemntary Figure S3

Chapter 6

IDEOM: An Excel interface for analysis of LC–MS based metabolomics data

Darren J. Creek[1,2], Andris Jankevics[3,4], Karl E. V. Burgess[1], Rainer Breitling[3,4], Michael P. Barrett[1]

Bioinformatics 28(7):1048–1049, 2012

1 Institute of Infection, Immunity and Inflammation, Wellcome Trust Centre for Molecular Parasitology, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, United Kingdom
2 Department of Biochemistry and Molecular Biology, Bio21 Molecular Science and Biotechnology Institute, University of Melbourne, Melbourne, Australia
3 Institute of Molecular, Cell and Systems Biology, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, United Kingdom
4 Groningen Bioinformatics Centre, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Groningen, The Netherlands

Summary: *The application of emerging metabolomics technologies to the comprehensive investigation of cellular biochemistry has been limited by bottlenecks in data processing, particularly noise filtering and metabolite identification. IDEOM provides a user-friendly data processing application that automates filtering and identification of metabolite peaks, paying particular attention to common sources of noise and false identifications generated by liquid chromatography–mass spectrometry (LC–MS) platforms. Building on advanced processing tools such as mzMatch and XCMS, it allows users to run a comprehensive pipeline for data analysis and visualization from a graphical user interface within Microsoft Excel, a familiar program for most biological scientists.*

Availability and implementation: *IDEOM is provided free of charge at* `http://mzmatch.sourceforge.net/ideom.html`*, as a macro-enabled spreadsheet (.xlsb). Implementation requires Microsoft Excel (2007 or later). R is also required for full functionality.*

Contatc: `michael.barrett@glasgow.ac.uk`

Metabolomics aims to measure all small molecules (metabolites) in a biological system. However, challenges associated with data analysis, in particular the accurate identification of metabolites (Moco *et al.*, 2007; Neumann and Böcker, 2010) have restricted progress.

Recent advances in high resolution mass spectrometry (MS) provide accurate mass detection that significantly improves metabolite identification from liquid chromatography–MS (LC–MS) (Breitling *et al.*, 2006b; Moco *et al.*, 2007) although limitations still abound (Kind and Fiehn, 2010; Neumann and Böcker, 2010), and noise or artifact peaks are often incorrectly identified as metabolites (Scheltema *et al.*, 2009).

Applications are freely available for detection, quantification and alignment of signals in LC–MS data (Blekherman *et al.*, 2011), and numerous multivariate statistical methods have been proposed to extract significant features from these high-dimensional datasets (Madsen *et al.*, 2010). However, many of the most powerful tools are implemented in statistical software, and while the graphical user interface (GUI) in MZmine (Pluskal *et al.*, 2010b) simplifies data pre-processing, recent advances in noise filtering and identification algorithms (Brown *et al.*, 2011; Scheltema *et al.*, 2011) are difficult to use without special training.

IDEOM is a Microsoft Excel template with a collection of VBA macros that enable automated data processing of high resolution LC–MS data from untargeted

metabolomics studies, with a particular focus on removal of noise [which accounts for ∼80% of peaks in typical LC–MS metabolomics datasets (Jankevics *et al.*, 2012)], metabolite identification and data visualization. Its GUI allows users to exploit the power of data-processing methods such as mzMatch (Scheltema *et al.*, 2011) and XCMS (Smith *et al.*, 2006) from within Excel, and enables rapid and simple conversion of raw or pre-processed LC–MS data into a filtered, interpretable list of putative metabolites, with associated confidence levels and sample intensities.

6.1 Methods

Opening the IDEOM template in Excel provides a spreadsheet on which to specify the parameters for data processing (default values are provided). All IDEOM macros for data processing are then activated by buttons or in-cell hyperlinks, and pop-up dialog boxes are used to specify important parameters for individual processing steps.

Raw LC–MS data files (.mzXML) are processed by IDEOM using the freely available XCMS (Smith *et al.*, 2006) and mzmatch.R tools (`http://mzmatch.sourceforge.net/index.php`) (Scheltema *et al.*, 2011), by automatic generation and execution of scripts in the R environment (`www.r-project.org`) (R Core Team, 2012) using readily adjustable parameters. Raw peaks are extracted by XCMS, and mzMatch applies peak matching, noise filtering, gap-filling and annotation of related peaks. Pre-processed peak lists from mzMatch or MZmine can be directly imported into IDEOM as text or csv files. During data import, any number of samples can be assigned to (up to 30) study groups (e.g. blanks, controls and QCs) to improve data filtering, analysis and visualization. Sample consistency, according to average peak height and internal standards (optional), is automatically checked and normalization can be applied.

Noise filtering within IDEOM removes common sources of noise in high resolution LC–MS data: chromatographic peak shoulders, irreproducible peaks, background or contaminant signals, and fourier transform (FT) and electrospray ionisation (ESI) artifacts including isotopes, adducts and fragments (based on Brown *et al.* (2011) and Scheltema *et al.* (2011)), as described in the documentation (Supplementary Material).

Metabolite identification is achieved by matching the accurate mass and retention time of observed peaks to metabolites in the included database, which incorporates all likely metabolites from a wide range of biological databases, and can be updated by users for specific applications. Retention times for authentic standards, and a retention time prediction model, are included for ZIC-HILIC chromatography data

(Creek *et al.*, 2011); however, as retention times are instrument specific, users are encouraged to upload standard retention times from their own platform with the macro provided. Subsequent to initial metabolite identification, a data-dependent polynomial mass recalibration step can be applied to correct for mass-dependent calibration errors. The final lists of identified, and rejected, peaks are annotated with confidence scores regarding the identification of each metabolite, based on retention times, organism-specific (or user-defined) databases and annotation as possible 'related peaks' (Scheltema *et al.*, 2009). Univariate statistics (mean, relative intensity, SD, t-test and Fisher ratio) are calculated in Excel. Multivariate statistics are obtained by calls to the R environment.

6.2 Results

The automated pre-processing steps in IDEOM drastically reduce the need for manual curation of LC–MS data by applying filters to remove hundreds of false-identifications (Creek *et al.*, 2011). The example tutorial dataset (Supplementary Material) demonstrates putative identification of 1314 metabolites and rejection of 1942 false-identifications from serum samples. The putatively identified metabolites are displayed in an interactive table that includes metabolite information, sample intensities, comparative statistics, LC–MS data, and links to websites and graphs (Fig. 6.1). Putative identifications can be rapidly scrutinized by double-click access to LC–MS metadata, or adjusted by selecting alternative isomers from dropdown lists. Data visualization is enhanced by Excel's conditional formatting, filtering and sorting, allowing users to view results according to sample intensities, significance, pathways or other properties.

Additional IDEOM tools include: merging dual-polarity data, chemical formula determination for unidentified masses, stable isotope tracking and targeted analysis. Full documentation, tutorials, source code and the IDEOM template are freely available at `http://mzmatch.sourceforge.net/ideom.html`.

IDEOM provides a user-friendly interface for analysis of complex metabolomics datasets without the need for specialist bioinformatics skills, allowing for the rapid production of meaningful, interactive results for biological interpretation of untargeted metabolomics data sets.

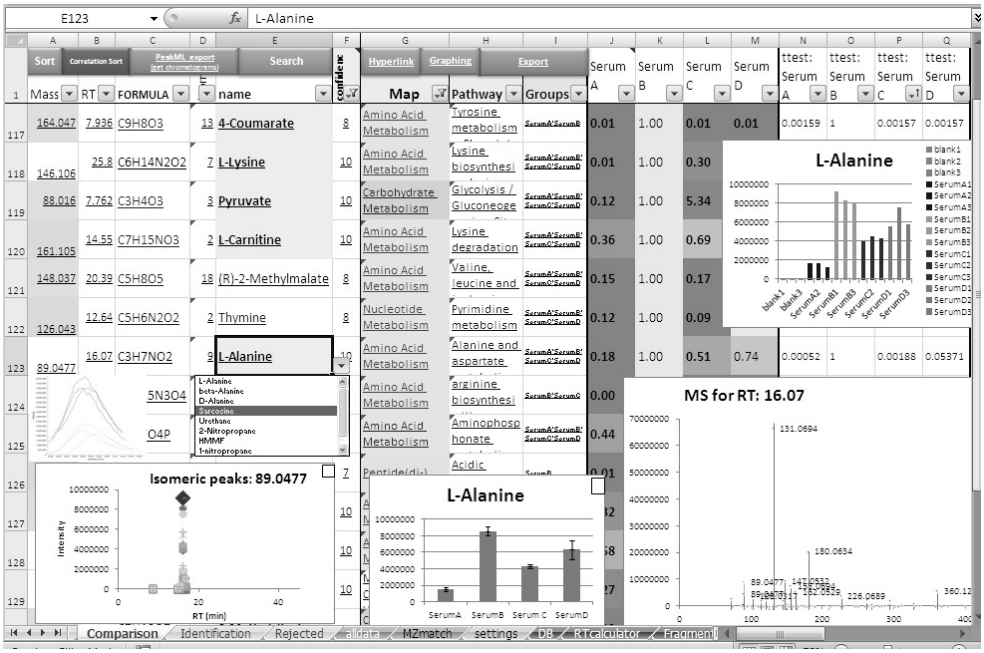


Figure 6.1: Screenshot of results visualization showing the dropdown list selection of isomers and some of the automatically generated hyperlinked charts.

Acknowledgements

The authors thank Gavin Blackburn, Isabel Vincent and Anubhav Srivastava for testing and feedback. Example datasets were provided by the Scottish Metabolomics Facility, funded by the Scottish Universities Life Sciences Alliance (SULSA).

Funding: Australian National Health and Medical Research Council postdoctoral training fellowship. Funding for AJ was provided by an Netherlands Organisation for Scientific Research NWO-VIDI grant to RB.

Conflict of Interest: none declared.

Supplementary Material

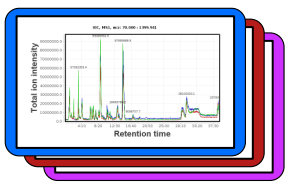
Additional information as noted in text. This material is available free of charge via the Internet at `http://bioinformatics.oxfordjournals.org/content/28/7/1048/suppl/DC1`.

Chapter 7

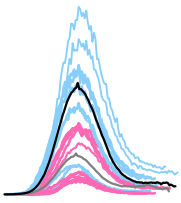
MzMatch/mzMatch.R: an open source software for the sequential processing and analysis of mass spectrometry data – A tutorial

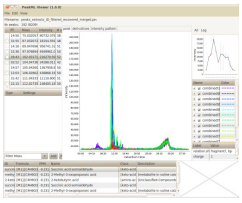
Andris Jankevics

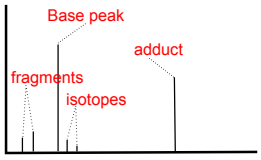
http://mzmatch.sourceforge.net



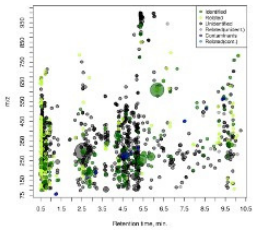
Peak detection, advanced filtering

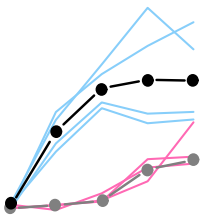




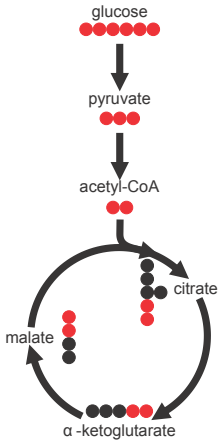


Matching of related peaks





Identification of detected peaks, visualisation and statistics



Stable isotope labeling

abstract

In this tutorial a customizable pipeline for mass spectra calibration, processing, filtering, annotation, statistical analysis and visualization is presented. The benefits of such a processing pipeline include an easy "rewind" option to roll back to intermediate steps in the data analysis and increased verifiability of the performance of the analytical methods. The developed approaches have provided us with a better understanding of the information content of our observations and a better assessment of the metabolites detected in the analyzed data sets.
A working example of the generic mzMatch processing pipeline is given, applying the software to an LC/MS data set.

7.1 Prerequisites

Download and install R version 2.15.3 or later (`http://www.r-project.org/`). R is an integrated suite of software facilities for data manipulation, calculation and graphical display. If you are unfamiliar with the R syntax and commands, it is recommended to read the "Introduction to R" document first (`http://cran.r-project.org/doc/manuals/r-release/R-intro.html`).

Download and install Oracle Java 6 JRE (Java Runtime Environment) (`http://www.oracle.com/technetwork/java/javase/downloads/index.html`) or later. On machines running 64-bit Windows operating systems, both the 32-bit and the 64-bit Java version must be installed.

The mzmatch and XCMS programs are supporting data files in mzXML, mzML and mzData format. For converting between manufacturers' proprietary formats and these supported data formats we recommend to use the "msconvert" tool from the ProteoWizard package (`http://proteowizard.sourceforge.net/formats/index.html`).

The mzmatch.R package depends on a number of other libraries, which need to be installed on your R- system. Once R has been installed, the other packages required can be installed by pasting the following into the R GUI Console:

```
source("http://bioconductor.org/biocLite.R")
biocLite(c("xcms", "multtest", "mzR"))
install.packages(c("rJava", "XML", "snow", "caTools",
   "bitops", "ptw", "gplots", "tcltk"))
```

As a last step, the mzmatch.R library need to be installed, calling an R script:

```
source ("http://puma.ibls.gla.ac.uk/mzmatch.R/install_mzmatch.R")
```

7.1.1 A very short guide to the R syntax

Most of the operations carried out in R involve the evaluation of functions. Functions are invoked by their name with a list or parameters separated by commas. For example:

```
PeakML.Viewer (arch="detect", install.path=NULL, JHeapSize=1024,
   uninstall=FALSE)
```

In this case, the `PeakML.Viewer` function has four parameters, which are assigned to default values that the user can override when needed. The set of parameters and their default values are set in the function definition and usually described in detail in the on-line documentation.

To execute a function using all default parameters you should enter a function name and opening and closing brackets:

```
PeakML.Viewer ()
```

To adjust a single argument and use the rest with default values, the function call should be like this:

```
PeakML.Viewer (install.path="D:/PeakMLView")
```

In this case we changed a location where the PeakML Viewer package should be installed or is located.

To store the output of a function as an object in R memory, we can use assignments:

```
FolderName <- getwd ()
```

In this example, an object ”FolderName” will store the full name of the current project directory (output from the getwd function).

In a similar way, also numerical or character strings can be assigned:

```
Mass <- 234.34095
SampleName <- "Sample1"
```

7.1.2 PeakML Viewer

A user interface application called PeakML Viewer enables quick visualisation, inspection and manipulation of the contents of a PeakML file (e.g. manual selection and/or export of peaks of interest). In this tutorial, screen shots for the output of some data processing steps are provided. More information about the tool can be found in **Chapter 2, Section 2.2.2** and on `http://mzmatch.sourceforge.net/peakmlviewer.php`. To install and run the PeakML Viewer you can use the following commands within R:

```
require(mzmatch.R)
PeakML.Viewer ()
```

7.1.3 Documentation

Documentation on all functions shown in this tutorial can be accessed from the mzmatch website (`http://mzmatch.sourceforge.net/mzmatch.R/00Index.html`) or in the R on-line help system.

```
help.start ()
```

After the R documentation page is opened in the Web browser, navigate to "Packages ⇒ mzmatch.R".

7.2 Data processing

7.2.1 File structure and data processing setup

Establishing an easy-to-use and maintainable file structure is key when processing any large data sets. All data files should be stored on the local hard drive and not on shared network drives or remote file systems.

Nineteen data files recorded on an Orbitrap Exactive mass spectrometer coupled to an HPLC chromatography system will be used in this tutorial. Data files were kindly provided by the Unit of Molecular Parasitology of the Institute of Tropical Medicine, Antwerp, Belgium (Dr. Maya Berg and Dr. Manu Vanaerschot).

To access this data set you can use following commands in R; we assume that you have a read/write access to your "D:" drive(on machine running Microsoft Windows). If not, you will have to alter the file paths accordingly.

```
setwd ("D:/")
download.file("http://puma.ibls.gla.ac.uk/datafiles/tutorial_data.zip",
   destfile="tutorial_data.zip")
unzip ("tutorial_data.zip")
file.remove ("tutorial_data.zip")
```

In case the direct download link does not work or access to a Web resources from R is restricted by a proxy server you can download the data set used in this tutorial from here: `http://mzmatch.sourceforge.net/mzmatch.tutorial.data.html`. After downloading the file "tutorial_data.zip", uncompress the file and follow the further steps of this tutorial.

filenames	**sampleClass**	**globalClass**
./mzXML/Blank.mzXML	Blank	Blank
./mzXML/Class_1A.mzXML	Class_1	Data
./mzXML/Class_1B.mzXML	Class_1	Data
./mzXML/Class_2A.mzXML	Class_2	Data
./mzXML/Class_2B.mzXML	Class_2	Data
./mzXML/DilS_1A.mzXML	DilS_1	QC
./mzXML/DilS_1B.mzXML	DilS_1	QC
./mzXML/DilS_2A.mzXML	DilS_2	QC
./mzXML/DilS_2B.mzXML	DilS_2	QC
./mzXML/DilS_4A.mzXML	DilS_4	QC
./mzXML/DilS_4B.mzXML	DilS_4	QC
./mzXML/DilS_8A.mzXML	DilS_8	QC
./mzXML/DilS_8B.mzXML	DilS_8	QC
./mzXML/STD1_A.mzXML	STD_1	Standards
./mzXML/STD1_B.mzXML	STD_1	Standards
./mzXML/STD2_A.mzXML	STD_2	Standards
./mzXML/STD2_B.mzXML	STD_2	Standards
./mzXML/STD3_A.mzXML	STD_3	Standards
./mzXML/STD3_B.mzXML	STD_3	Standards

Table 7.1: An example of a data processing setup file. The first column indicates the name and location of the data file. In the second column the sample class of each individual data file is defined. The third (optional) column defines a sample grouping, if the data set can be divided in more global classes than provided by the sampleClass definition.

In either case, the folder "tutorial_data" will be created. It contains the subfolder "mzXML", where the actual data files converted to the mzXML file format are stored. The file "sample_setup.tsv", formated as tab separated values (TXT), contains information about technical/biological replicates, and defines a type and sample class for each sample. The file should have at least two columns and the first row should contain the labels. The following columns are expected: "filenames"

– a relative path (if files are located in subfolders the file name should start with "./folder_name") and the full name of the data files. "sampleClass" – a character string defining a class label for the data files. "globalClass" – this column is optional and can contain a label to group together different sample sub-classes, for example, analytical samples and quality control (QC) samples.

Contents of the "sample_setup.tsv" file for the tutorial data set are shown in Table 7.1. The data set contains one "Blank" sample (measurement extraction solvent); this sample/samples can be used to filter out biologically non-relevant signals from the data set. Samples in groups "Class_1" and "Class_2" are 2 biological replicates measured for two different biological conditions. Samples from groups "DilS_1", "DilS_2", "DilS_4" and "DilS_8" are dilution series samples which will be used to filter out spurious signals as described in **Chapter 4**. Samples "STD_1", "STD_2" and "STD_3" are measurements of mixtures of authentic standards (**Chapter 3, Section 3.1.1**) and can be used for metabolite identification in the analytical samples.

Before the mzmatch.R package can be used, it needs to be loaded into the R environment. This can be achieved with the following command:

```
library(mzmatch.R)
mzmatch.init(version.1=FALSE)
```

NOTE: This tutorial is based on mzmatch.R 2.0 (the function parameter version.1 is set to FALSE). PeakML files produced with older versions of mzmatch.R will not be compatible with mzmatch.R 2.0. Functions included in version 2 are backward compatible with the syntax of old commands, but some tools mentioned in this tutorial are not supported in versions below 2.0.

After the mzmatch.R library is initialised, we can run a command to set up the data analysis. If the file "sample_setup.tsv" has already been created the following command should be used:

```
mzmatch.R.Setup(samplelist="sample_setup.tsv")
```

This command will open a graphical window (Figure 7.1) asking the user to navigate to the project folder "D:/tutorial_data" (file "sample_setup.tsv" should be located at the root of this folder). Note that all data files can be stored in one or more sub-folders of this project folder.

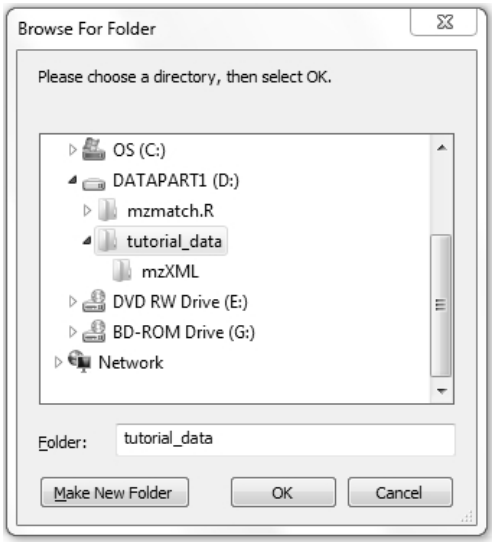


Figure 7.1: Graphical window asking to setup a main project folder.

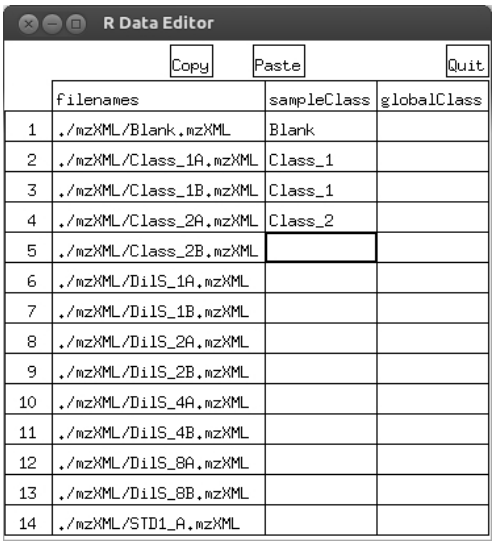
Alternatively, you can specify a "projectFolder" parameter inside the function, indicating a project directory. This parameter is useful if you want to run data processing in the batch mode or on remote servers, where interactive input is not possible.

```
mzmatch.R.Setup(projectFolder="D:/tutorial_data",
   samplelist="sample_setup.tsv")
```

Contents of the "sample_setup.tsv" files can be generated also directly within R, calling a function without any parameters.

```
mzmatch.R.Setup ()
```

After selecting a data folder from the interactive menu window, all files with extensions mzXML, mzData or mzML present in a given folder and its sub-folders will be listed and the user will be asked to enter "sampleClass" values (Figure 7.2). After finishing entering all data and clicking on "Quit", the data will be stored and the file "sample_setup.tsv" created in the project folder (if you are using Microsoft Windows just close the editor window to save the data).



	filenames	sampleClass	globalClass
1	./mzXML/Blank.mzXML	Blank	
2	./mzXML/Class_1A.mzXML	Class_1	
3	./mzXML/Class_1B.mzXML	Class_1	
4	./mzXML/Class_2A.mzXML	Class_2	
5	./mzXML/Class_2B.mzXML		
6	./mzXML/DilS_1A.mzXML		
7	./mzXML/DilS_1B.mzXML		
8	./mzXML/DilS_2A.mzXML		
9	./mzXML/DilS_2B.mzXML		
10	./mzXML/DilS_4A.mzXML		
11	./mzXML/DilS_4B.mzXML		
12	./mzXML/DilS_8A.mzXML		
13	./mzXML/DilS_8B.mzXML		
14	./mzXML/STD1_A.mzXML		

Figure 7.2: Interactive window to edit sample data from the `mzmatch.R.Setup()` command call.

7.2.2 Extracting peaks with the centWave algorithm from XCMS

In **Chapter 1 and 2** we already discussed the general data processing steps. The first step in data analysis is feature or peak detection. The command below will use centWave (a highly sensitive features algorithm developed for use with high resolution LC–MS data) algorithm (Tautenhahn *et al.*, 2008) from the XCMS library to extract peaks from raw data files. The parameter "nSlaves" defines how many parallel processes are run simultaneously, so that multiple cores or processors can be used for faster data analysis.

```
xseto <- xcmsSet(sampleList$filenames, method='centWave', ppm=2,
   peakwidth=c(10,100), snthresh=5, prefilter=c(3,1000),
   integrate=1, mzdiff=0.01, verbose.columns=TRUE,
   fitgauss=FALSE, nSlaves=4)
```

After xcms set is created, in the next step we would be exporting the detected signals to the PeakML file format. The command below will create a single file for each measurement, and by default output files are placed in the folder "peakml" within the project directory. 19 files will be created; each of these files can be readily used in the PeakML Viewer.

```
PeakML.xcms.write.SingleMeasurement(xset=xseto,
```

```
filename=sampleList$outputfilenames, ionisation="detect", ppm=5,
   addscans=0, ApodisationFilter=TRUE, nSlaves=4)
```

7.2.3 Combine biological replicates

At this step, also called matching, detected signals are aligned between technical or biological replicates of a single sample. This file combination step allows applying different filters (noise filter, RSD filter etc.) on peak data. As replicated samples are expected to be most similar to each other, and any variability in signal intensities is likely to be related to the stability of measurement or sample degradation, such filters are very effective to eliminate irreproducible and noisy peaks from the data sets.

The unit of the "rtwindow" parameter is seconds, and it defines how similar the retention time must be for a peak between different data files for it to be considered as the same signal. The "ppm" parameter is used to set a window of mass deviation in ppm unit. These should be determined by the performance of the instrument that the samples were analysed on.

Output files from the command are written to a folder "combined" in project folder. The folder will contain 10 output files, the number of sample classes defined during the experiment setup step (see Table 7.1 column "sampleClass").

```
mzmatch.ipeak.Combine(sampleList=sampleList, v=T, rtwindow=60,
   combination="set", ppm=5, nSlaves=4)
```

7.2.4 RSD filtering

In this example, the RSD filter (Scheltema *et al.*, 2008), to remove the most variable (i.e. irreproducible) signals between replicates, is applied on the combined data sets. Two folders in the tutorial data directory are created. In the folder "combined_RSD_filtered" data sets of filtered files are written, while in the folder "combined_RSD_rejected" rejected mass chromatograms are written. Note that, for example, we have only one sample in the group "Blank", so that it is not possible to calculate RSD values of the signal intensity for this sample. In this case the RSD filter function will not discard any signals in the given sample.

```
mzmatch.ipeak.filter.RSDFilter(sampleList=sampleList, rsd=0.3,
   v=T, nSlaves=4)
```

7.2.5 Combine by conditions

After removal of spurious and irreproducible signals within replicates of the single samples we can apply a combiner function, to again combine together RSD filtered data files in one final data file. In this way, data from all experimental conditions are grouped together, providing the possibility to compare results and apply further filtering steps on the data set. A single file named "final_combined.peakml" is created.

```
INPUTDIR <- "combined_RSD_filtered"
FILESf <- dir (INPUTDIR,full.names=TRUE,pattern="\\.peakml$")
mzmatch.ipeak.Combine(i=paste(FILESf,collapse=","), v=T, rtwindow=60,
   o="final_combined.peakml", combination="set", ppm=5)
```

As was mentioned already above, the PeakML.Viewer software can be used to explore the contents of every generated PeakML file. For example, we can check a peaks in the latest output file.

```
PeakML.Viewer ()
```

After the PeakML Viewer program is loaded, we can use the "File ⇒ Open" menu to open any PeakML file from the project folder. A screenshot from the file "final_combined.peakml" is shown in Figure 7.3. One can see that for mass 90.054912 a significant retention time (RT) drift is present.

To solve issues related to RT drifts, we can try to apply a correction method provided with the XCMS package. In this example, the Obiwarp (Prince and Marcotte, 2006) method will be applied. We will repeat the same commands as we used in example above, but with slightly different function parameters. The command `retcor` is used to apply RT correction on the XCMS set object created before. Then we change the `mzmatch.R setup` function call to output Peakml files to a new location (parameter: "outputfolder"), so that a comparison between corrected and uncorrected data is possible. In a similar way we adjust the call for combiner and RSD filter functions. PeakML files for single measurements are created in the "peakml_RTcorr" folder, and combined files of replicates in "combined_RTcorr", and and RSD filter output in "combined_RTcorr_RSD_filtered". The final combined file will be named "final_RT_corr_combined.peakml".

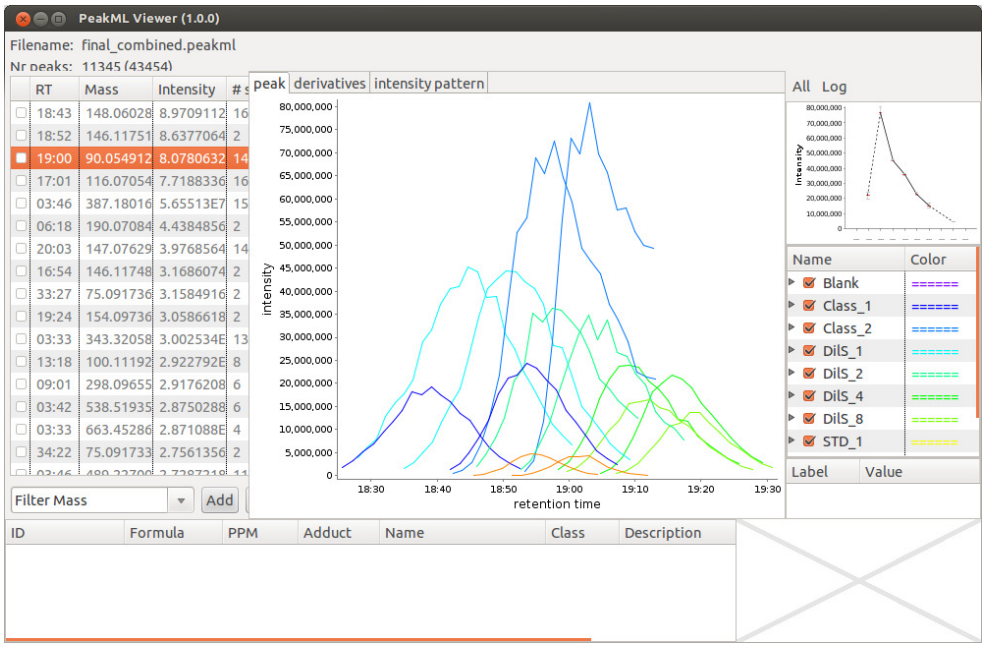


Figure 7.3: Combined data set visualized in the PeakML Viewer program. The highlighted peak is showing significant drift in retention time.

```
mzmatch.R.Setup(projectFolder=getwd(), samplelist="sample_setup.tsv",
   outputfolder="peakml_RTcorr")
xset2<-retcor(xseto, method="obiwarp", profStep=0.01, center=3)
PeakML.xcms.write.SingleMeasurement(xset=xset2,
   filename=sampleList$outputfilenames, ionisation="detect", ppm=5,
   addscans=0, ApodisationFilter=TRUE, nSlaves=4)
mzmatch.ipeak.Combine(sampleList=sampleList, v=T, rtwindow=60,
   combination="set", ppm=5, nSlaves=4,outputfolder="combined_RTcorr")
mzmatch.ipeak.filter.RSDFilter(sampleList=sampleList, rsd=0.3, v=T,
   nSlaves=4, inputfolder="combined_RTcorr")
INPUTDIR <- "combined_RTcorr_RSD_filtered"
FILESf <- dir (INPUTDIR,full.names=TRUE,pattern="\\.peakml$")
mzmatch.ipeak.Combine(i=paste(FILESf,collapse=","), v=T, rtwindow=60,
    o="final_RT_corr_combined.peakml", combination="set", ppm=5)
```

A screen shot showing the same peak set after RT correction is given in Figure 7.4, and a considerable improvement in RT alignment can been seen.

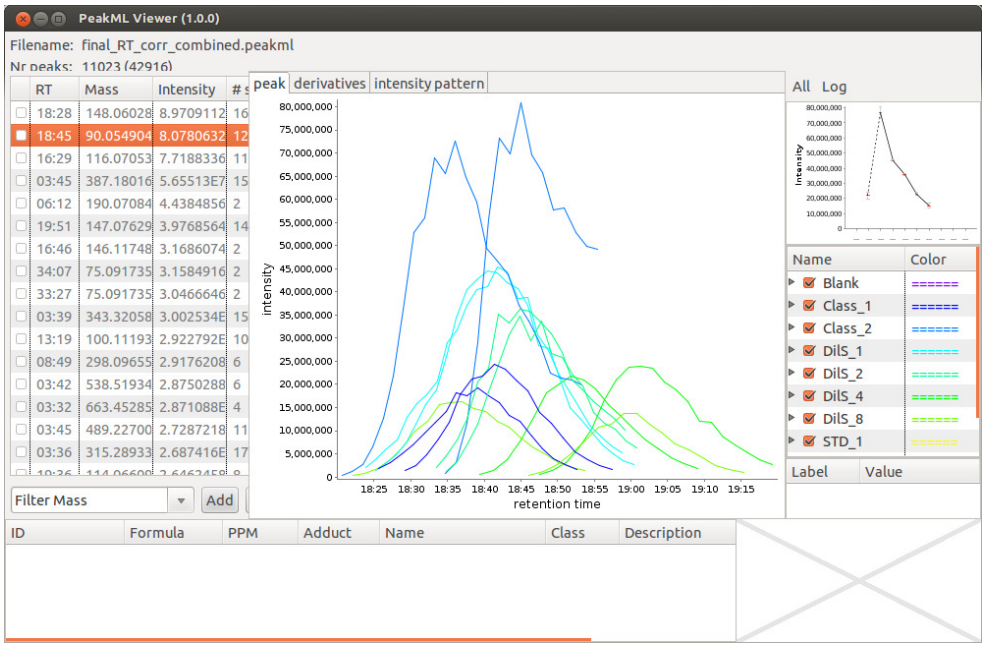


Figure 7.4: Combined data set visualized in the PeakML Viewer program after RT correction. Although RT drifts are still present, a significant improvement can be observed (compare to Fig. 7.3).

For further analysis we will use the data files after RT correction.

7.2.6 Blank filter

This filter compares signal intensities between analytical samples and blanks. If peak intensity in blanks is higher or equal to that measured in "real" samples, then these signals are discarded from the data set. The command below will create a file "final_combined_blank_filtered.peakml", and discarded chromatograms will be stored in the file "blank_removed_final_combined_blank_filtered.peakml". The parameter "BlankSample" is used to define which of the sample sets present in the PeakML file should be used as a blank sample.

```
PeakML.BlankFilter (filename="final_RT_corr_combined.peakml",
   ionisation="detect",
   outputfile="final_combined_blank_filtered.peakml",
   BlankSample="Blank")
```

7.2.7 Gap filler tool

It is possible that the combiner incorrectly missed signals, or signals were missed by the peak picking algorithm because of the irregular shape of signals of low intensity. After samples from different conditions are grouped together, it is essential to estimate the background signal intensity level (if a peak was discarded by XCMS because of the peak shape) or if the signal is really absent in the current case. This tool complements missing signals from the raw data files to a combined file produced in the previous step. Also, in Figure 7.4 it is clearly shown that highest intense peaks are partially cut at 18 min 50 sec. If the parameter "fillAll" is set to "TRUE", then also signals present in the input file will be re-integrated from the raw data files, using the full RT and mass window of the measured peak set.

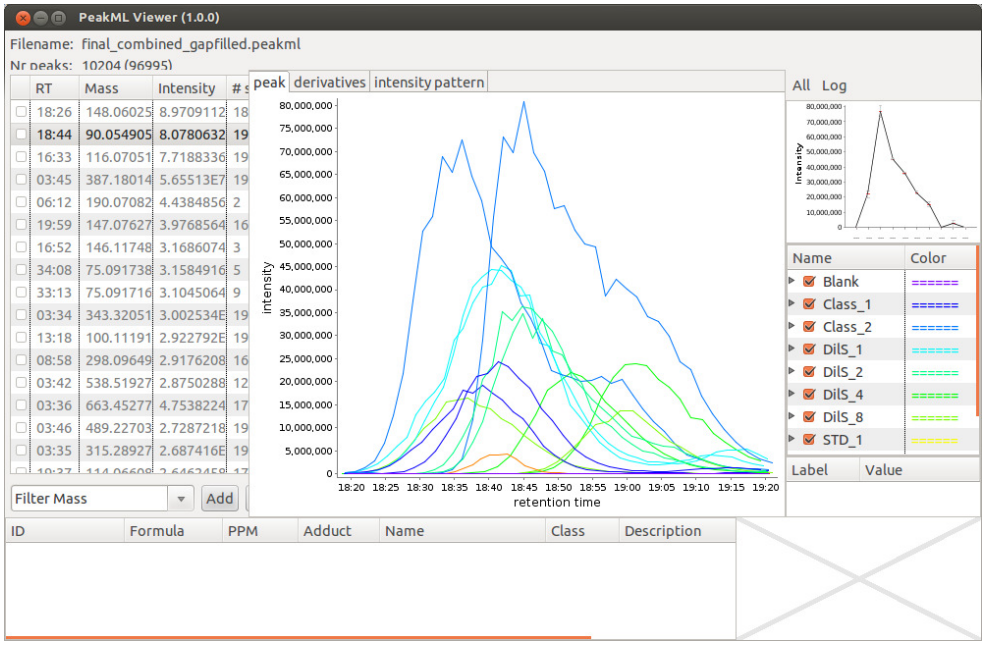


Figure 7.5: Data set after application of the gap filling tool. Missing or incomplete peaks have been filled in, especially peak shoulders and low intensity signals (compare to Fig. 7.4).

A file named "final_combined_gapfilled.peakml" is created in the tutorial working directory.

```
PeakML.GapFiller(filename = "final_combined_blank_filtered.peakml",
   ionisation = "detect",
```

```
   outputfile = "final_combined_gapfilled.peakml", ppm = 0,
   rtwin = 0, fillAll=TRUE, Rawpath=NULL)
```

In Figure 7.5, an example output of the Gap filler tool is shown.

7.2.8 Simple filter

The output of the Gap filler tool generated a file with 10204 peak sets, but some peak sets still do not contain peaks for all 19 samples; this means that in some samples there was no signal detected within the current mass and RT range of these peak sets. Usually, signals detected in only a very small fraction of the all samples contains noise or irrelevant information (this can depend on the type of study, but often the same metabolites are expected to be present in all samples, only at different levels). The following command is retaining only those peak sets which contain peaks from at least 6 samples. The `mzmatch.ipeak.filter.SimpleFilter` function can be also used to filter out signals with intensity lower than some selected threshold, specific mass or RT window or specific annotations. Please refer to the on-line documentation for more details.

A file named "final_combined_ndetect.peakml" is created.

```
mzmatch.ipeak.filter.SimpleFilter( i="final_combined_gapfilled.peakml",
   o="final_combined_ndetect.peakml", mindetections=6, v=T)
```

7.2.9 Dilution trend filter

To remove uninformative signals we will also apply a dilution trend filter as described in **Chapter 4**. The parameter "trendSets" defines sample set names of samples from a dilution series sorted from the most concentrated to the most diluted. All signals that do not correlate with the expected dilution trend (p-value larger than 0.10) will be discarded.

A file named "final_combined_DF.peakml" is created. Discarded peak sets are stored in the file "discarded_final_combined_DF.peakml".

```
trendSets <- c("DilS_1", "DilS_2", "DilS_4", "DilS_8")
PeakML.DilutionTrendFilter (filename="final_combined_ndetect.peakml",
   ionisation="detect", Rawpath=NULL, trendSets=trendSets,
   p.value.thr=0.10, outputfile="final_combined_DF.peakml")
```

7.2.10 Match related peaks

This tool clusters groups of related peaks (e.g., isotope peaks, adducts, fragments, multiply charged molecules) together and attempts to identify their relationship. Each cluster is given a unique id, which is stored as an annotation with the name "relation.id". The relationship is stored as an annotation with the name "relation.ship". The possibility is offered only to store the most intense peak of each cluster (option "basepeak"). This is useful for cleaning up the data before attempting statistical approaches to mine the data. However, as the peak of interest is not per definition the most intense peak one can not rely on this file alone for identification purposes, but needs to apply careful judgment based on the total evidence.

Files named "final_combined_related.peakml" (for the full data set) and "final_combined_basepeaks.peakml" (for base peaks only) are created.

```
mzmatch.ipeak.sort.RelatedPeaks(i="final_combined_DF.peakml", v=T,
   o="final_combined_related.peakml",
   basepeaks="final_combined_basepeaks.peakml", ppm=5, rtwindow=30)
```

7.2.11 Identify peaks from databases

This function matches the contents of a given PeakML file with molecular formulas and identities of compounds in various databases. Within the tool chain, in the `package mzmatch.ipeak.db`, several tools are provided for converting downloadable files (usually from an ftp-server) from the major metabolite databases. When a peak is positively identified, the annotation "identification" is added with the unique database ID corresponding to the match.

Nine example databases are bundled with the mzmatch.R. In the following example, the first command assigns a list of available databases to the R variable "DBS". The second command outputs this list of databases on the screen. The third command creates a comma separated character string with database files that we want to use for compound identification. In this example we are using the ESI contaminants database (Keller *et al.*, 2008) (number 4 in the list), a database of *Leishmania donovani* parasite metabolites (number 5), and three standard mixture database files (numbers 1 to 3).

By default, the identification tool will match a measured mass with a compound mass from a data base taking into account that depending on the electrospray source ionisation mode M+H or M−H ions will be detected. The parameter "adducts" can be used to specify a list of additional adducts (mass differences) to consider in the search. In the example below, for positive ionisation mode data we use the following

list of adducts: M+H, M+ACN+Na, M+Na, M+K, M+ACN+H. In Figure 7.6, the output of the identification tool is shown; all details concerning peak identification are shown in the bottom panel of the PeakML Viewer.

Output file named "final_combined_related_identified.peakml" (for complete peak file) is created.

```
DBS <- dir(paste(.find.package("mzmatch.R"), "/dbs", sep=""),
   full.names=TRUE)
DBS
DBS <- paste(DBS[c(1,2,3,4,5)],collapse=",")
mzmatch.ipeak.util.Identify(i="final_combined_related.peakml", v=T,
   o="final_combined_related_identified.peakml", ppm=5, databases=DBS,
   adducts="M+H,M+ACN+Na,M+Na,M+K,M+ACN+H")
```

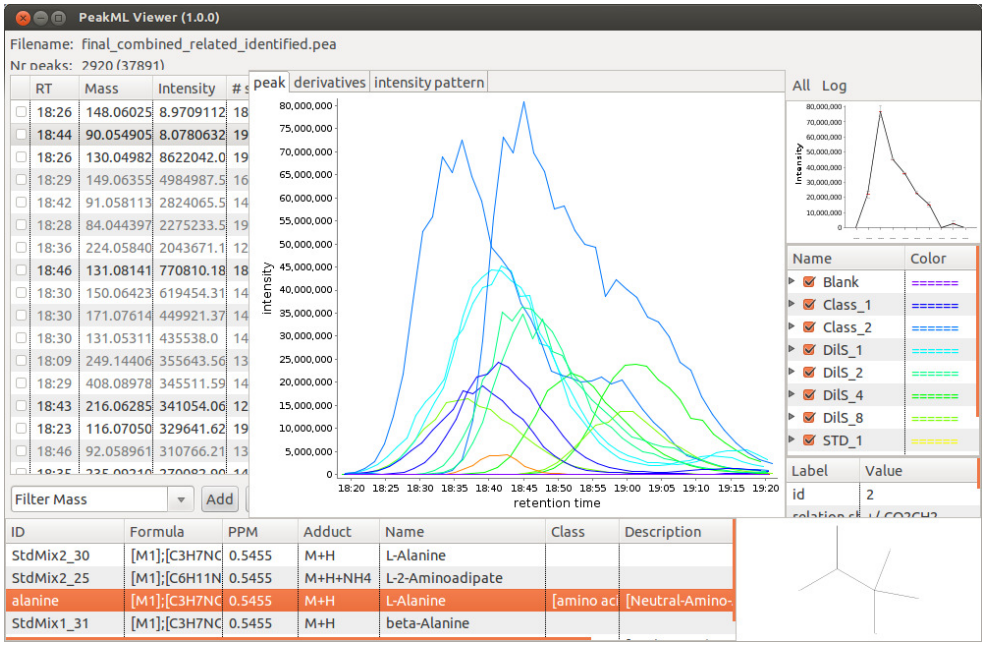


Figure 7.6: Information about identified peaks is added to the PeakML file, displayed in the bottom pane of the PeakML Viewer. Note that, where available, the structural formula of the molecules is also displayed at the bottom right.

7.2.12 Convert to text

This function produces a TAB separated TXT file from the final PeakML file produced by the process. The TXT file can be opened in either notepad, Excel, or imported in tools like IDEOM (**Chapter 6**). If opened in Excel, the first column will be the m/z for the peak and the second will be the retention time. The next few columns will contain the intensity information for the compound in each sample, and the final column will contain the putative identification. This allows you to further process the data if you require.

Output file named "final_combined_related_identified.txt" (for complete peak file) is created.

```
mzmatch.ipeak.convert.ConvertToText (
   i="final_combined_related_identified.peakml",
   o= "final_combined_related_identified.txt", databases=DBS,
   annotations="identification,ppm,adduct,relation.ship")
```

7.2.13 Plot related peaks clusters

The function below will generate a graphical output in pdf file format. It takes related peaks clusters assigned by the `mzmatch.ipeak.sort.RelatedPeaks` command and plots an overview, sorted by peak intensity. The fisrt plot (Figure 7.7) of each grouped related peaks set is showing a global mass spectrum, including the average signal intensity across all samples at a specified retention time. Peaks colored in red are signals containing putatively assigned identifications. The following plots show extracted peak chromatograms for each sample, peak mass, retention time, intensity and putatively assigned identifications.

Output file named "Related_peaks_plot.pdf" is created.

```
PeakML.Plot.RelatedPeaks (
   filename="final_combined_related_identified.peakml",
   ionisation="detect", Rawpath=NULL, DBS=unlist(strsplit(DBS,",")),
  outputfile="Related_peaks_plot.pdf",
   sampleClasses=c("Class_1","Class_2","STD_1","STD_2","STD_3"))
```

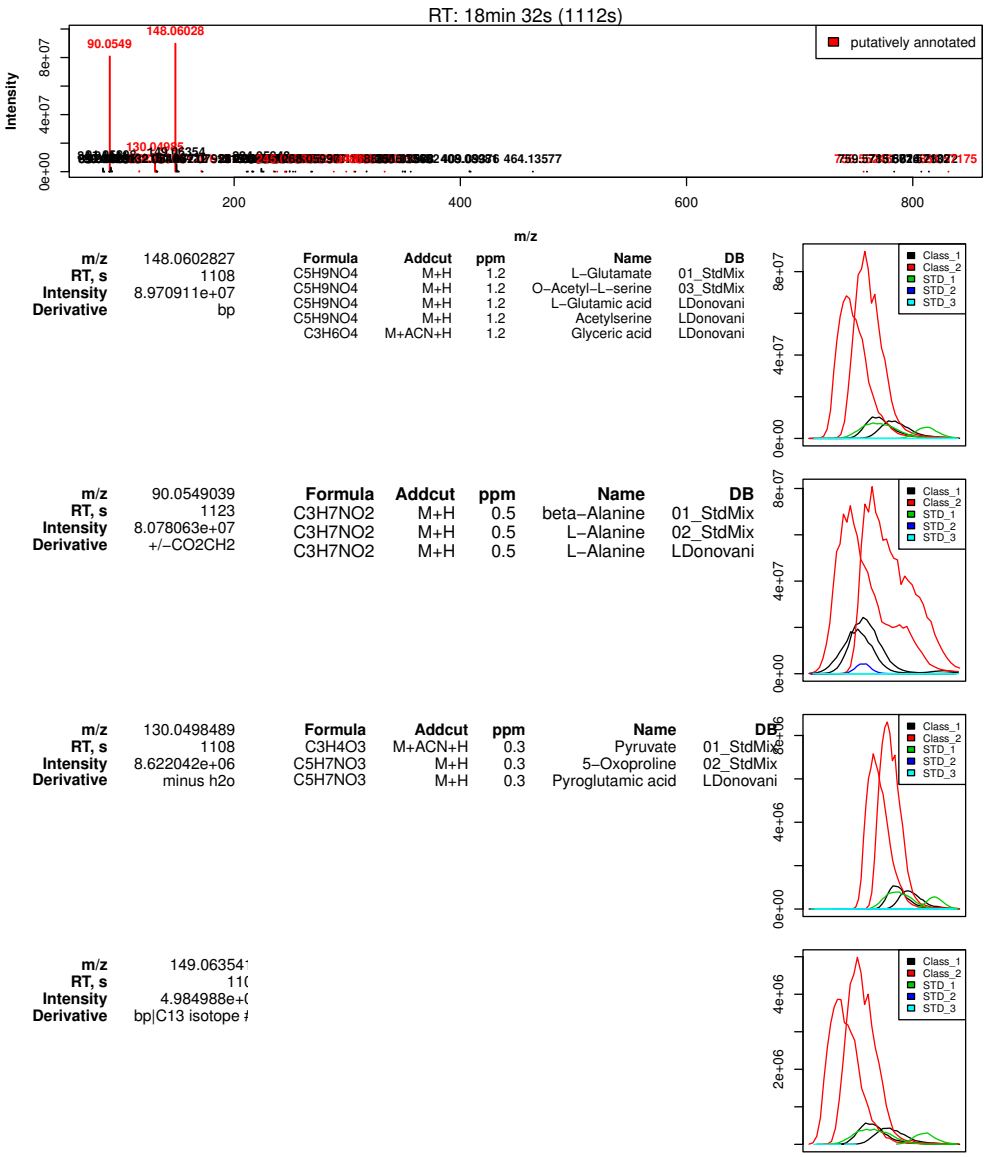


Figure 7.7: Output generated by the PeakML.Plot.RelatedPeaks command. Two signals with high intensity are putatively identified, and the rest of the signals show much lower intensity. It is most probable that the measured mass 148.06003 can be annotated as L-Glutamine, while the mass 90.05490 is a glutamine fragment that is misidentified as L-Alanine (note that mzMatch identifies this putative relationship between the peaks, indicating the possible loss of CO_2CH_2). The water loss and first C^{13} isotope peak are also correctly identified).

7.3 Advanced Processing

As we discussed in **Chapter 2**, one main advantage of mzmatch and the PeakML file format is the possibility to integrate the final or intermediate results with other tools. The examples below will give you a short tutorial on accessing data from R for visualization or statistical purposes.

To read a PeakML data file in the R environment, the `PeakML.Read` function is used. Output of the function is an R PeakML object, which contains the complete information stored in the original PeakML file, including sample names, extracted chromatograms, sample group names and annotations, peak identifications etc. For more details on function parameters and output structure please refer to the help pages.

```
PeakMLData <- PeakML.Read("final_combined_related_identified.peakml")
```

This command creates an R object "PeakMLData". Each measured peak in this object is represented as a chromatogram. In **Section 7.2.12** we gave an example how to export peak intensity information in TXT file format. Although it is possible to read these output files back into the R environment, below we show an example how to get the same structure also directly from a "PeakMLData" object.

```
PeakTable <- PeakML.Methods.getCompleteTable (PeakMLData)
```

The resulting object is a list of three data tables (matrices). The number of rows in each data table is equal to the number of samples, and the number of columns is equal to the number of peak sets (or peak groups) in the PeakML file. The first table contains peak intensities data, the second table masses, and the third retention times.

Such 3-dimensional data for each single sample can be visualised for example as a "bubble plot" (Figure 7.8), where on the x-axis the measured RT is plotted, on the y-axis the mass, and the decimal logarithm of signal intensity determines the size of each "bubble". Such plots are an effective and quick way to evaluate the quality of chromatographic separation and the mass/retention time regions where most of the signals are detected. The example code below will create a pdf file "Bubble_plots.pdf", in which a bubble plot for each sample will be plotted on a single page. The sample names will be used as plot title.

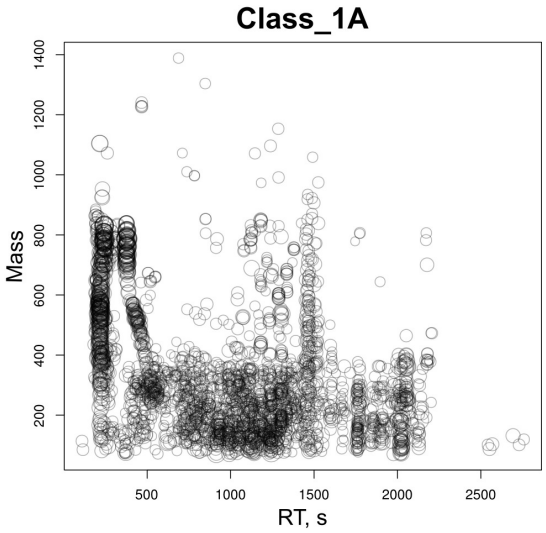


Figure 7.8: Example output of the "bubble plot". It can be concluded that large quantity of masses with m/z larger than 400 are eluting at the beginning of the chromatogram. This data set was collected using HILIC chromatography, and it is clearly shown that such setup is not useful to separate lipids and proteins (with masses above m/z 400). A large fraction of small molecules (including also related peaks) are ealuting between 8 and 25 (480 – 1500 s) minutes of chromatography.

```
pdf ("Bubble_plots.pdf")
for (snum in 1:length(PeakMLData$phenoData)) {
   plot (PeakTable$Retentiontimes[snum,], PeakTable$Masses[snum,],
      pch=21, cex=log(PeakTable$Intensities[snum,])/5,
      xlab="RT, s", ylab="Mass")
      title(main=PeakMLData$sampleNames[snum])
}
dev.off ()
```

In the next example we will show how to apply a t-test on selected samples from the data. Of course, as there are only 2 samples in each group, one would usually not use a t-test on this dataset, but the same function call would be applied to datasets with a larger number of replicates.

To compare "Class_1" with "Class_2", we need to determine which rows in the data table are matching samples of interest for each sample class:

```
class1_hits <- which(PeakMLData$phenoData=="Class_1")
class2_hits <- which(PeakMLData$phenoData=="Class_2")
```

Now we can calculate a t-test for each of the peak sets and store the p-values. Note that if in one or both of the measurements, no signal was detected, then it is not possible to apply a t-test to this peak set. In that case, in the example below, such peak sets will be skipped. After the t-test, the Benjamini and Hochberg method is

used to adjust calculated p-values for multiple comparisons (Benjamini and Hochberg, 1995).

```
pvalues <- rep(NA,ncol(PeakTable[[1]]))
for (pval in 1:length(pvalues)) {
   data_class1 <- PeakTable$Intensities[class1_hits,pval]
   data_class2 <- PeakTable$Intensities[class2_hits,pval]
   NA1 <- length(which(is.na(data_class1)))
   NA2 <- length(which(is.na(data_class2)))
   if ( NA1==0 & NA2==0) {
      pvalues[pval] <-t.test(data_class1, data_class2)$p.value
   }
}
adjusted.pval <- p.adjust (pvalues, method="BH")
```

The resulting p-values can be added as an extra annotation to the "PeakMLData" object and written back to a PeakML file. Afterwards, the resulting file can be opened in the PeakML Viewer, where for example peaks can be selected by setting a p-value threshold.

```
PeakMLData$GroupAnnotations$p.values.adjusted <- adjusted.pval
PeakML.Write (PeakMLData,outFileName="processed.peakml")
```

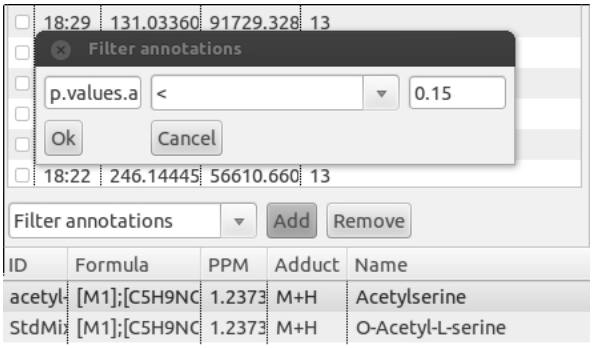


Figure 7.9: An example how to filter peak sets in the PeakML Viewer based on annotation values.

To filter a PeakML file by annotation in PeakML Viewer, locate the drop-down menu on the left side at the bottom of PeakML Viewer program. Select the "Filter annotations" from the list and click the "Add" button. A small configuration window will be opened (Figure 7.9). In the first field, enter the annotation name (in our example it is "p.values.adjusted"). In the second field, operators equal (=), larger(>)

or smaller(<) should be entered. In the third field a threshold for the value can be set, in the current example we selected 0.15. After pressing the "OK" button, 49 peak set matching these criteria will be displayed in the PeakML Viewer. If you want to filter peaks based on whether an annotation is present for given peak set, then in the first field you should enter the annotation name (for example, "identification") and in the second field the symbol ">", leaving the third field empty and pressing "OK". The PeakML Viewer will show only peak sets containing an "identification" annotation (see **Chapter 2, Section 2.2.2**).

Alternatively, of course, peak sets of interest can be visualised directly within the R environment. For example, to create a pdf file, which contains all peak sets with p-value smaller than 0.15, we first have to select a peak sets matching this criterion, and then construct a plotting command.

```
matched.peaks <- which(adjusted.pval <= 0.15)
```

We want to add averaged mass across all samples of the peak set and the p-value from the t-test, as a plot title:

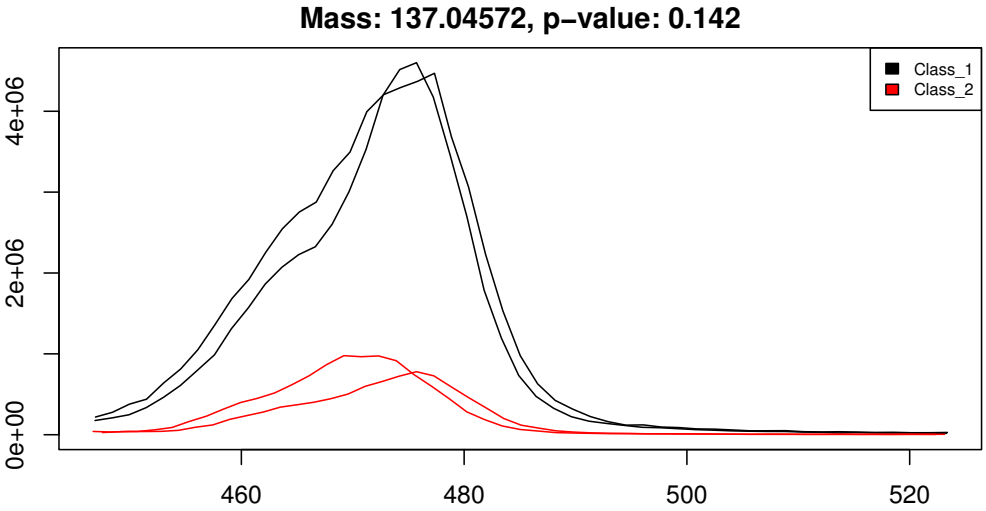
```
Masses <- apply(PeakTable[[2]], 2, mean, na.rm=TRUE)
titles <- paste ("Mass: ", round(Masses,5), ", p-value: ",
   round(adjusted.pval,3), sep="")
```

And add putative identifications if those are present for current data set:

```
DBS <- dir(paste(.find.package("mzmatch.R"), "/dbs", sep=""),
   full.names=TRUE)
DBS <- DBS[c(1,2,3,4,5)]
id.resolved <- PeakML.Methods.DBidToCompoundName(DBS=DBS,
   PeakMLdata=PeakMLData, collapse=FALSE)
```

Finally we can create a plotting routine; in this case, a pdf file "pvalues_plot.pdf" will be created, each page of the document contains a chromatogram plot of the selected peak set and peak identification information if it is available (Figure 7.10).

```
pdf ("pvalues_plot.pdf")
for (peak in matched.peaks) {
   par(mfrow=c(2,1))
   PeakML.Plot.Chromatograms (PeakMLdata=PeakMLData, groupid=peak,
      sampleClasses=c("Class_1","Class_2"), xaxis=TRUE,
      Title=titles[peak], mar=c(2,2,2,2))
   textplot(id.resolved[[peak]], halign="left", valign="top",
      mar=c(1,1,1,1))
}
dev.off ()
```



	id	formula	mass	name	db	ppm
105	StdMix2_1	C5H4N4O	136.0385	Hypoxanthine	02_StdMix	0.6
1326	hypoxanthine	C5H4N4O	136.0385	Hypoxanthine	LDonovani	0.6

Figure 7.10: An example output of the PeakML.Plot.Chromatograms function used to generate a visual report on selected peak sets.

7.4 Summary

All commands shown in this tutorial can be used as a single R script, and the complete example script can be located in the tutorial data set and is called "Tutorial.R". When first learning to use R and mzMatch to analyze your data, the main point is to realize that each step in the script happens successively and creates a new file. This means

that if you insert a new function somewhere in your processing, you must make sure that the function that occurs after this one is looking for the right input file. The functions used to process the data have been selected to produce a set of results that should accurately show the metabolites detected within the set parameters. This is a basic set up that uses fairly strict filters to produce an initial data set for the user to evaluate. The parameters can be changed in the script and this will alter the output of the processing.

To run the whole script open it via the file menu in R. Now run this script via the edit menu, "Edit ⇒ Run all" or you can use the `source` command from the R console.

Chapter 8

Towards an unbiased metabolic profiling of protozoan parasites: optimisation of a *Leishmania* sampling protocol for HILIC–Orbitrap analysis

Ruben t'Kindt[1,4], Andris Jankevics[2,3], Richard A. Scheltema[2], Liang Zheng[4], David G. Watson[4]. Jean-Claude Dujardin[1], Rainer Breitling[2,3], Graham H. Coombs[4] and Saskia Decuypere[1]

Anal. Bioanal. Chem. 398(5):2059–2069, 2010

1 Department of Parasitology, Unit of Molecular Parasitology, Institute of Tropical Medicine, Antwerp, Belgium
2 Groningen Bioinformatics Centre, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Groningen, The Netherlands
3 Faculty of Biomedical and Life Sciences, University of Glasgow, Glasgow, United Kingdom
4 Strathclyde Institute of Pharmacy and Biomedical Sciences, University of Strathclyde, Glasgow, United Kingdom

abstract

Comparative metabolomics of Leishmania species requires the simultaneous identification and quantification of a large number of intracellular metabolites. Here, we describe the optimisation of a comprehensive metabolite extraction protocol for Leishmania parasites and the subsequent optimisation of the analytical approach, consisting of hydrophilic interaction liquid chromatography coupled to LTQ-Orbitrap mass spectrometry. The final optimised protocol starts with a rapid quenching of parasite cells to 0 °C, followed by a triplicate washing step in phosphate-buffered saline. The intracellular metabolome of 4 × 10^7 parasites is then extracted in cold chloroform/methanol/water 20/60/20 (v/v/v) for 1 h at 4 °C, resulting in both cell disruption and comprehensive metabolite dissolution. Our developed metabolomics platform can detect approximately 20% of the predicted Leishmania metabolome in a single experiment in positive and negative ionisation mode.

Leishmania is a group of kinetoplastid protozoans that infect mammals following transmission by sand flies (*Psychodidae: Phlebotominae*). The life cycle of this parasite consists of two distinct morphological forms: (a) promastigotes, adapted to life in the sand fly intestinal tract and easily cultured axenically *in vitro*, and (b) amastigotes, living intracellularly within mammalian macrophages. *Leishmania* parasites cause a wide spectrum of poverty-related neglected diseases called leishmaniasis, a disease characterised by diversity and heterogeneity endemic in large areas of the tropics, subtropics and the Mediterranean basin. The heterogeneity of *Leishmania* species is reflected in the variety of clinical manifestations they cause, ranging from self-healing skin lesions to disfiguring mucocutaneous lesions up to lethal visceral disease. Even infection with one particular *Leishmania* species can result in diverse clinical profiles in terms of disease severity and treatment response. This clinical polymorphism challenges health professionals to manage leishmaniasis patients effectively (Chappuis *et al.*, 2007). Unfortunately, little is known about the intra- and inter-species molecular differences underlying this clinical polymorphism (Scheltema *et al.*, 2010; Smith *et al.*, 2007). Therefore, characterising the diversity of parasite populations is considered as a key step towards a better control of this disease. Metabolomics is an emerging field that allows in-depth characterisation of the metabolome, the closest correlate to the phenotype on molecular level (Faijes *et al.*, 2007; Scheltema *et al.*, 2010). This technology has great potential for resolving the metabolic variation between different *Leishmania* species and among multiple parasite isolates of a single species (reviewed in Scheltema *et al.* (2010)).

Protocols for monitoring the metabolome of protozoan species have not been fully optimised yet, in contrast to other unicellular organisms (Faijes *et al.*, 2007; Meyer *et al.*, 2010; Winder *et al.*, 2008). This study aimed to optimise a sample preparation protocol for subsequent liquid chromatography–mass spectrometry-based metabolomics analysis of *Leishmania donovani* promastigotes.

Generally, a sampling protocol needs to start with immediate arrest of the intracellular metabolism. Several methods are described in the literature for unicellular organisms, ranging from immersing the cells in cold organic solvent followed by centrifugation of the quenched cells (Faijes *et al.*, 2007; Villas-Bôas and Bruheim, 2007) to applying rapid filtration of cells followed by freezing them with liquid nitrogen (Bolten *et al.*, 2007; Meyer *et al.*, 2010). It is essential that, during this initial step, rupture of the cell envelope causing leakage of intracellular metabolites is prevented. Furthermore, removal of the extracellular environment is a challenging task, especially since *L. donovani* is generally cultured in rich medium with 20% (*v/v*) foetal calf serum. De Souza *et al.* (2006) and Saunders *et al.* (2010) have described rapid (seconds) quenching of *Leishmania* cultures to 0 °C using a dry ice–ethanol bath. They reported that this method quenches metabolism of *Leishmania* reliably and reproducibly.

In the second step, the quenched and washed cells need to be disrupted. Chemical and physical dissimilarities between cell envelopes hamper the design of a universal cell disruption method for global metabolomics analysis of single-cell organisms. Some species simply burst by permeabilisation in organic solvents (e.g. *Lactobacillus plantarum* (Faijes *et al.*, 2007), *Escherichia coli* (Maharjan and Ferenci, 2003) or cultured mammalian cells (Sellick *et al.*, 2009), while others demand mechanical interference to disrupt the rigid cell envelope (e.g. *Chlamydomonas reinhardtii* (Bölling and Fiehn, 2005; Lee and Fiehn, 2008) or *Mycobacterium bovis* (Jaki *et al.*, 2006)). *Leishmania* promastigotes are protected by a specific membrane enclosure, i.e. a thick glycoconjugate surface coat also referred to as the glycocalyx (Naderer and McConville, 2008). Since adequate cell lysis is essential to release the bulk of intracellular metabolites, we evaluated six different cell disruption procedures based on heating, milling or mixing of promastigotes.

Finally, the released metabolites need to be characterised. Hydrophilic interaction liquid chromatography (HILIC) coupled to LTQ-Orbitrap mass spectrometry (liquid chromatography–mass spectrometry, LC–MS) has already been identified as an excellent tool for metabolic profiling of *Trypanosome* parasites, a distant relative of *Leishmania* (Kamleh *et al.*, 2008). Phospholipids tend to be retained with reversed-phase chromatography (C18), causing ion suppression effects over long periods of

chromatography by co-eluting into the ion source with other analytes of interest (Annesley, 2003; Kamleh *et al.*, 2009a). However, with HILIC chromatography, these lipid compounds elute in the void volume, allowing more reliable quantification of compounds eluting later. These polar metabolites, which encompass the majority of the intracellular metabolome, include signature metabolites of the trypanosomatid species such as trypanothione. Moreover, LC–MS covers a wide mass range which allows analysis of many compound classes not detectable by the alternative gas chromatography coupled to MS (Böttcher *et al.*, 2007).

Here, we report the optimisation of a complete protocol for metabolome analysis of *Leishmania*, from the quenching of promastigotes, cell disruption and metabolite extraction, up to the HILIC–Orbitrap analysis. The full protocol was validated by applying it to three *L. donovani* isolates.

8.1 Experimental

8.1.1 Chemicals and materials

Formic acid (ULC grade), acetonitrile (ULC grade), water (ULC grade), methanol (ULC grade) and chloroform (high-performance liquid chromatography (HPLC)-S grade) were purchased from Biosolve (Valkenswaard, the Netherlands). The ZIC®-HILIC PEEK Fitting Guard columns and ZIC®-HILIC PEEK HPLC columns were obtained from HiChrom (Reading, UK). Phosphate-buffered saline (PBS) was obtained from Invitrogen (Merelbeke, Belgium).

Regarding the culturing of *Leishmania* promastigotes, modified Eagle's medium (designated HOMEM medium, Invitrogen) (Mottram *et al.*, 1992) has been supplemented with 20% (*v/v*) heat-inactivated foetal calf serum (PAA Laboratories GmbH, Linz, Austria) pH 7.5 at 26 °C. Disposable count chambers (Uriglass) were obtained from Menarini Diagnostics (Reading, UK). The *L. donovani* strains MHOM/NP/02/BPK282/0 and MHOM/NP/03/BPK275/0 were isolated from bone marrow aspirates from confirmed visceral leishmaniasis patients recruited at the B.P. Koirala Institute of Health Sciences, Dharan, Nepal, as described by Rijal *et al.* (2007). The clinical parasite isolates were identified as *L. donovani* based on a CPB PCR–RFLP assay (Tintaya *et al.*, 2004). The *L. donovani* strain 1S has its origin in Sudan (MHOM/SD/00/1S-2D). Extraction/cell disruption devices included Thermomixer (Eppendorf AG, Hamburg, Germany), Retsch mill (Retsch GmbH & Co. KG, Haan, Germany), Ultra Turrax mixer (IKA Werke GmbH & Co KG, Staufen, Germany) and Dispomix (Xiril AG, Hombrechtikon, Switzerland).

High-performance liquid chromatography equipment consisted of a Surveyor HPLC pump (ThermoElectron, Hemel Hempstead, UK) and two different ZIC-HILIC column setups differing mainly in the column diameter: either (a) a ZIC-HILIC guard column (20 mm × 2.1 mm; 5 μm) and analytical column (150 mm × 4.6 mm; 3.5 μm) or (b) a ZIC-HILIC guard column (15 mm × 1.0 mm; 5 μm) and analytical column (150 mm × 2.1 mm; 3.5 μm). High-resolution mass measurements were obtained with a Finnigan LTQ-Orbitrap mass spectrometer (Thermo Fisher Scientific Inc., Hemel Hempstead, UK). The LC–MS system was controlled by Xcalibur version 2.0 (Thermo Fisher Scientific Inc., Hemel Hempstead, UK).

8.1.2 Parasite growth conditions

Leishmania promastigotes were grown in modified Eagle's medium (Mottram *et al.*, 1992) supplemented with 20% (*v/v*) heat-inactivated foetal calf serum pH 7.5 at 26 °C. The cultures were initiated by inoculating parasites from a culture at days 3–4 of stationary phase of growth into 5 ml culture medium to a final concentration of 5×10^5 parasites per millilitre. Independently growing cultures of the parasite were treated as biological replicates.

After 7 days of growth, when parasites had been in stationary phase for 4 days, parasite density was determined using disposable count chambers, and subsequently an aliquot corresponding to 1×10^8 parasites (for optimisation cell disruption/metabolite extraction experiments) or 4×10^7 parasites (for all other experiments) was taken from each culture. Experiments are described following their order in the final sample preparation protocol. However, the quenching method was evaluated after optimisation of the cell disruption/metabolite extraction step and the LC–MS approach. All optimisation experiments were done with BPK282/0 in late-stationary promastigote stage, except for the cell leakage part, where both BPK282/0 and 1S were used.

8.1.3 Metabolite extraction

Optimisation experiments for cell disruption and metabolite extraction were performed in triplicate (biological replicates), using one-phase chloroform/methanol/water 20:60:20 (*v/v/v*) at 0 °C as a comprehensive extraction solvent. Following harvest of 1×10^8 promastigotes (BPK282/0 isolate) and three wash steps in PBS, cell disruption procedures were applied. Tested procedures

include heating block (HB; 70 °C; 60 min), Thermomixer (TM; 60 min; 4 °C; 1,400 rpm), vortex with glass beads (V; 0.5 mm i.d.; 3 × 1 min; cool in between in ice bath), Retsch mill with glass beads (RM; 0.5 mm i.d.; 3 × 1 min; cool in between in ice bath), Ultra Turrax mixer (UT; 3 × 1 min; in ice bath) and Dispomix (DM; 3 × 1 min; 0 °C). After cell disruption and metabolite extraction, all samples were centrifuged for 10 min at 16,100 × g (4 °C). The resulting supernatants were separated from cell debris and analysed immediately on the HILIC–Orbitrap platform. The raw data files of all different extraction procedures are processed and aligned together.

The cell leakage experiment was done with two *L. donovani* strains, BPK282/0 and 1S. Three biological replicates were grown for each strain and harvested as described above. Their spent culture medium (M) and the successive washing solutions (W1-3) were aliquoted during the sample preparation and stored at –70 °C until further analysis within the next 48 h. Before LC–MS analysis, 75 μl of each fraction was taken and deproteinised with 300 μl 20:60 chloroform/methanol (*v/v*) to obtain identical solubility of compounds as in the parasite extracts. Following centrifugation (16,100 × g; 4 °C; 10 min), deproteinised supernatant was separated and used for analysis.

The final, optimised metabolite extraction protocol consists of (a) quenching (<20 s) of *L. donovani* promastigotes in their culture flasks to 0 °C in a bath containing a mixture of dry ice/ethanol, (b) isolating the necessary volume for harvesting 4 × 10^7 parasites, (c) triplicate washing of parasite cells in 1 ml of cold (0 °C) PBS (pH 7.4) by centrifugation (20,800 × g, 0 °C, 3 min) and re-suspending cells using a vortex, (d) cell disruption and metabolite extraction of the washed cell pellet in 200 μl chloroform/methanol/water 20:60:20 (*v/v/v*) during 1 h in a Thermomixer (1,400 rpm, 4 °C), (e) isolating the metabolite extract from cell debris by centrifugation (20,800 × g, 0 °C, 3 min) and (f) deoxygenating the extracts with a gentle stream of nitrogen gas for 1 min prior to tube/vial closure (Fig. 8.7).

8.1.4 LC–MS analysis

Gradient elution was performed on two different ZIC-HILIC column setups differing mainly in the column diameter: either (a) 150 mm × 4.6 mm or (b) 150 mm × 2.1 mm. Elution of the ZIC-HILIC columns was carried out with a gradient of (A) 0.1% formic acid in acetonitrile and (B) 0.1% formic acid in water. The flow rate was 300 and 100 μl/min, with an injection volume of 10 and 5 μl, respectively. Gradient elution chromatography was always performed starting with 80% solvent A. Within a 6-min time interval, solvent B was increased to 40% and maintained for 12 min,

followed by an increase to 90% within 4 min. This composition was maintained for 2 min, after which the system returned to the initial solvent composition in 2 min. The whole system was allowed to re-equilibrate under these conditions for 14 min.

The LTQ-Orbitrap mass spectrometer was operated in both positive and negative ion electrospray mode. Optimal instrument parameters were based on previous results (Kamleh *et al.*, 2008, 2009a,b). Briefly, ESI source voltage was optimised to 4.0 kV, and capillary voltage was set to 30 V. The source temperature was 250 °C, and the sheath and auxiliary gas flow rates were 30 and 10, respectively, in machine-specific units. Full-scan spectra were acquired over an m/z range of 50–1,000 Da, with the mass resolution set to 30,000 full width at half maximum (FWHM). The target for mass accuracy was <1 ppm. By using a resolution of 30,000 FWHM, this was routinely achievable and allowed rapid spectrum acquisition compatible with the peak widths obtained by the chromatographic system, leading to at least 20 full scans across the width of a peak. All spectra were collected in continuous single MS mode.

8.1.5 Data processing

Raw data files acquired from analysed samples were converted to the mzXML format by the readw.exe utility (a tool of the Trans-Proteomic Pipeline software collection, downloaded from `http://tools.proteomecenter.org/wiki/index.php?title=Software:ReAdW`). Further processing was handled by a flexible data-processing pipeline mzMatch (Scheltema *et al.*, 2008) (`http://mzmatch.sourceforge.net`). mzMatch is a modular, open-source and platform-independent data-processing pipeline for metabolomics LC–MS data written in the Java language. It was designed to provide small tools for the common processing tasks for LC–MS data. The mzMatch environment was based entirely on the PeakML file format and core library, which provides a common framework for all the tools. mzMatch comprises integrated chemistry (e.g. molecular formulae, mass conversion and periodic table), math (e.g. statistics, wavelet transform, function fitting and loess and Savitzky-Golay) and visualisation (JFreeChart and SWT for user interface applications) routines. Starting from the PeakML data, signal detection (Olsen *et al.*, 2005), retention time alignment (Christin *et al.*, 2008; Windig, 2004), blank removal, noise removal (Windig, 2004) and signal matching were performed. Masses whose abundance was not reproducible for all biological and technical replicates, as indicated by a relative standard deviation (RSD) larger than 35%, were discarded, as quantification is expected to be at least 20% accurate over multiple runs (Shah *et al.*, 2000). Derivative signals (isotopes, adducts, dimers and

fragments) were automatically annotated by correlation analysis on both signal shape and intensity pattern (Scheltema *et al.*, 2009). The derivative signals were removed before further statistical tests, as they would give excessive weight to abundant analytes with many derivatives. The selected mass chromatograms were putatively identified by matching the masses (mass accuracy <1 ppm) progressively to those in metabolite-specific databases. In the first round of identification, LeishCyc (Doyle *et al.*, 2009), LipidMAPS (Fahy *et al.*, 2007) and a contaminant database were used (Keller *et al.*, 2008). The latter allows removal of typical impurities and buffer components often detected in metabolomics experiments. Only the remaining unidentified peak groups went through a second round of matching with KEGG (Ogata *et al.*, 1999) and a peptide database; and finally a third round was done with the Human Metabolome Database for any remaining unidentified analytes (Wishart *et al.*, 2007). This iterative process was used in order to restrict the number of potential matches to the most likely (Scalbert *et al.*, 2009). A selection of standards was run routinely to ensure that the instrument was returning true masses, thereby enabling the use of retention time in the identification process (Kamleh *et al.*, 2009b).

Statistical analysis and graphical routines were handled in R (R Core Team, 2012). Principal component analysis (PCA) is an unsupervised multivariate analysis technique frequently used in metabolomics (Trygg *et al.*, 2007). It implements a data dimensionality reduction of complex data matrices, so that clustering tendencies, trends and outliers can be visualised among samples. Unit variance scaling was used in PCA calculations, and no additional normalisation was performed on the datasets. The R code consisting of reading and writing routines of data from/to PeakML file format (XML representation of processed data produced by the mzMatch pipeline) is available from the authors upon request. Total metabolite yield is calculated as the sum of all metabolite intensities.

8.2 Results and discussion

8.2.1 Quenching and washing of *Leishmania* promastigotes

A quenching methodology based on De Souza *et al.* (2006) was assessed here for *L. donovani* promastigotes. The method involves quenching parasites in their culture medium, which avoids leakage of metabolites during quenching. This is in contrast with commonly used quenching methods based on suspension of cells in cold organic solvent, which increases the risk of cell lysis and subsequent loss of intracellular meta-

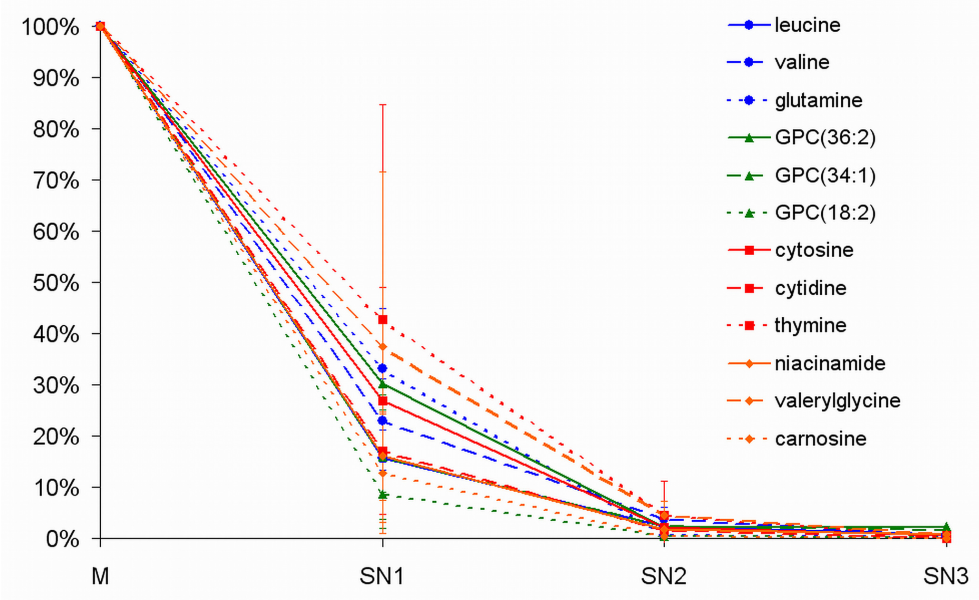


Figure 8.1: Washing efficiency–metabolite leakage during the successive washing steps for selected metabolites. Signal intensities in the SN of three sequential washing steps are compared to those of spent culture medium (M). The data clearly show that at least three washing steps are required to obtain clean parasite pellets without medium contamination (*GPC* glycerophosphocholine).

bolites (Lee and Fiehn, 2008; Meyer *et al.*, 2010; Wittmann *et al.*, 2004). However, in order to separate and extract only the intracellular *Leishmania* metabolome, it is crucial that the growth medium is effectively removed after the quenching step by washing the parasites thoroughly. These washing steps involve a risk of cell lysis and subsequent leakage of intracellular metabolites (Meyer *et al.*, 2010). PBS (pH 7.4), a common washing buffer for *Leishmania* promastigotes, was chosen here for the washing steps to remove culture medium after quenching. Possible leakage of cells was monitored by analysing the supernatants (SN) removed after each washing step using our LC–MS setup. Figure 8.1 shows the effective removal of selected medium compounds and extracellular metabolites during the successive washing steps. The data are representative for all detected metabolites. Phospholipids were the most difficult to remove; very small percentages (<2.35%) of the culture medium were still detected in the final wash solution. No intracellular metabolites (e.g. glutamate, glutathione, nicotinamide adenine dinucleotide (NAD) and trypanothione) could be detected in the supernatants of the different washing steps, confirming the absence

of leakage of intracellular metabolites. Overall, three successive washing steps are needed to remove the vast majority of culture medium compounds and extracellular metabolites.

8.2.2 Cell disruption and metabolite extraction

Comprehensive metabolite extraction is a crucial step in untargeted metabolomics. As no guideline currently exists on comprehensive metabolite extraction of *Leishmania* parasites, different cell lysis approaches were tested in order to disrupt the *Leishmania*-specific glycocalyx. All procedures were compared using chloroform/methanol/water 20:60:20 (*v/v/v*) as extraction solvent, which has been shown to provide comprehensive metabolite dissolution after cell disruption (Mal *et al.*, 2009). Procedures were based on heating (HB) (De Souza *et al.*, 2006; Kamleh *et al.*, 2008), mixing (Thermomixer, TM; Ultra Turrax, UT; Dispomix, DM), vortexing (vortex, V) or milling (Retsch mill, RM) (Jaki *et al.*, 2006). For each cell disruption procedure, three biological replicates were analysed. PCA clearly shows a denser clustering of samples prepared by TM or DM, compared to extracts prepared with the other approaches (Fig. 8.2), indicating a superior reproducibility with both methods. Calculation of the average RSD of the intensities of all detected preliminary annotated metabolites (a number of 63) confirms this (Table 8.1). The total metabolite yield of the detected metabolites was very similar for the different extraction procedures; only the RM method showed significantly lower signals.

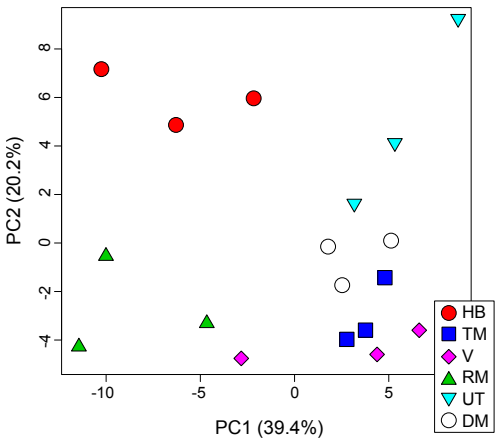


Figure 8.2: Principal component analysis of samples obtained by six different cell disruption/extraction procedures. The score plot representing the first two principal components explains 59.6% of the variation present in the LC–MS dataset. Thermomixer (TM) and Dispomix (DM) show the tightest clustering and thus the best reproducibility, compared with the other cell disruption methods (HB – heating block, V – vortex, RM – Retsch mill, UT – Ultra Turrax)

Some labile metabolites in the higher mass range, which are among the targeted compounds, showed some significant differences between the different methods. NAD

Method	Average RSD% ± SD	Total metabolite yield
HB	15.84%±6.71	1.73 $\times 10^8$
TM	9.81%±7.34	1.75 $\times 10^8$
V	15.91%±8.01	1.67 $\times 10^8$
RM	18.93%±7.17	1.35 $\times 10^8$
UT	17.97%±18.65	1.87 $\times 10^8$
DM	9.77%±7.94	1.78 $\times 10^8$

Table 8.1: Average relative standard deviation (RSD%) of the signal intensities and total metabolite yield of all annotated metabolites obtained with six different extraction procedures. RSD% was calculated for each metabolite detected (n = 3), and an average RSD%±SD was calculated for all metabolites. Total metabolite yield was calculated as the sum of all signal intensities. HB – heating block, TM – Thermomixer, V – vortex, RM – Retsch mill, UT – Ultra Turrax, DM – Dispomix

was detected in all cold approaches, while it was absent in the heated one (HB). On the other hand, nicotinamide could only be detected when heating was used, clearly showing up as a degradation product of NAD at elevated temperatures, as described before (Rover *et al.*, 1998). Trypanothione, a thiol metabolite typical for these parasitic protozoa (Fairlamb *et al.*, 1985), could not be detected in the heated method, while the level of its oxidised form, trypanothione disulfide, clearly exceeded the amount detected in the cold approaches. Ultra Turrax mixing of the parasites also led to a higher oxidation level of trypanothione. In conclusion, applying heat during metabolite extraction causes chemical degradation of some targeted metabolites.

Finally, the cold TM method was selected as the most adequate extraction approach for untargeted *Leishmania* metabolomics. This method was found to provide excellent reproducibility and good yield, and – in contrast to DM – it is an easy and fast method which is a relevant criterion for large-scale metabolomics studies.

Other extraction solvent compositions (aqueous methanol, aqueous ethanol, aqueous acetonitrile, aqueous isopropanol and 50:50 methanol/chloroform (*v/v*)) were tested alongside the chloroform/methanol/water 20:60:20 (*v/v/v*) one-phase composition. However, with the exception of 50:50 (*v/v*) methanol/chloroform, none of the alternative extraction solvents resulted in visual cell disruption when working in cold conditions (<4 °C). The presence of chloroform in the extraction solvent seems to be a key element for disrupting the *Leishmania* promastigote glycocalyx when using a cold-sample methodology. Other cell disruption methods might be advantageous with different solvents, but at the expense of reproducibility or degradation of labile metabolites.

Finally, the duration of the cell disruption process and extraction was also optimised. We compared extraction times of 15, 30 and 60 min for the selected Thermo-

mixer method for three technical replicates. The average relative standard deviation was 13.16%, 8.23% and 4.96%, showing that longer extraction times, somewhat surprisingly, led to more reproducible results. More specific experimental analysis should be done to identify possible reasons for this phenomenon.

8.2.3 Optimisation of cell number and analytical sensitivity

Initial experiments were done with 1×10^8 promastigotes in 1 ml of extraction solvent. *In vitro* cultures of *L. donovani* strains typically reach a density of $2.5 - 5 \times 10^7$ cells per millilitre in late-stationary phase. This means that a relatively large volume of cell culture (2–4 ml) is required to obtain 1×10^8 parasites, and this is unpractical for high-throughput experiments. Hence, we decreased the number of parasites to 4×10^7 cells, while at the same time decreasing the amount of extraction solvent to 200 μl, thereby increasing the actual cells/extraction solvent ratio by a factor of 2. This allowed handling of the parasite cultures in 1.5-ml tubes which further facilitated the extraction approach. The resulting metabolite profiles of three biological replicates showed an increased reproducibility (average RSD of 10.05%±7.40 for the 4×10^7 cells approach versus 13.09%±7.11 for the 1×10^8 cells approach) and a higher signal intensity for polar metabolites representing different metabolite classes (total signal intensity of 40 representative compounds is 1.9-fold higher with the 4×10^7 cells approach compared to the 1×10^8 cells approach).

HILIC was chosen as the optimal separation step, following the conclusions of earlier studies (Kamleh *et al.*, 2008). However, we further compared the performance of columns with different inner diameters, 4.6 mm (conventional HPLC as used in the previous studies) and 2.1 mm (micro-bore HPLC), to optimise analytical sensitivity (Fig. 8.3). Down-scaling the column diameter generally offers better detection sensitivity of metabolites (Granger *et al.*, 2005). This is confirmed by our results; a total of 43,580 signals were detected in positive ESI mode on the 2.1-mm column, compared with 28,060 signals on the conventional column. If we subtract signals originating from eluent and extraction solvent (Windig, 2004) and filter for reproducibility (35% RSD cutoff; (Shah *et al.*, 2000)) and good peak shape (Coda-DW value>0.8; (Windig, 2004)), the number of "quality peaks" in both analyses is 3,650 and 2,264, respectively, for the 2.1- and 4.6-mm column. A tentative identification based on database matching at a mass accuracy <1 ppm results in 390 versus 220 putative identifications for the 2.1- and 4.6-mm column. Thus, using a micro-bore column instead of a conventional HPLC column significantly increases the number of "quality peaks", resulting in an almost twofold increase in the number of putative identifications.

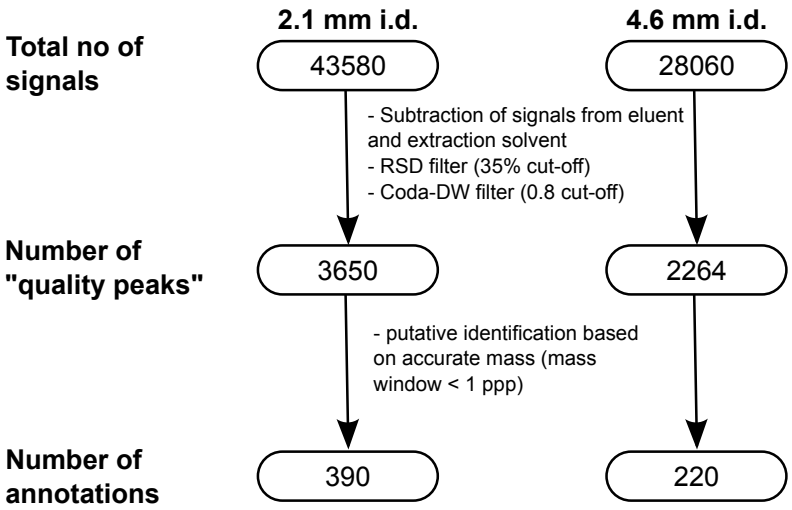


Figure 8.3: Comparison of the performance of columns with inner diameter 4.6 and 2.1 mm. The process of signal detection and signal filtering is described in the "Experimental" section.

8.2.4 Length of analytical block of LC–MS analysis of *Leishmania* extracts

The length of the analytical block can influence the quality of the analytical metabolomics data acquired with LC–MS platforms (Zelena *et al.*, 2009). Setting up the instrumentation is rather time-consuming, so everyone strives to analyse as many samples as possible in a single analysis batch.

However, instability of samples during autosampler storage can affect the metabolomics data negatively, showing decreasing trends of metabolites due to oxidation and/or degradation. Apart from this, the performance of the LC–MS analysis itself can decrease over time, leading to a decrease in sensitivity of the mass spectrometer, loss of mass accuracy and/or shifting of retention times. To determine the optimal analysis time, we analysed two identical extracts (designated A and B, aliquots taken from a pool of several technical replicates) repeatedly during 72 h to identify changes during autosampler storage of extracts. Extract B was placed in the autosampler 12 h after the analysis of extract A was started, allowing us to discriminate between a decrease in LC–MS performance and actual sample instability in the autosampler.

A decrease in signal of both extracts (A and B) at the same time would suggest a drop in LC–MS performance, while sample instability would show as an independent

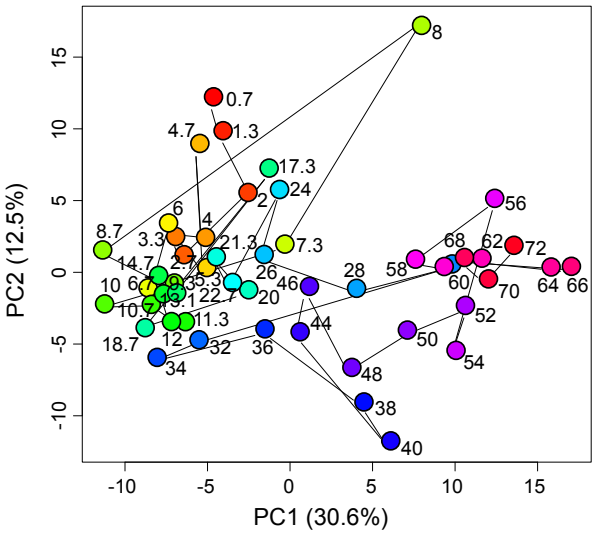


Figure 8.4: Principal component analysis showing the LC–MS time trend during analysis of extract A. The hour of analysis is indicated underneath each point in the plot, and points are connected in order of analysis. A constant drift in performance is visible in the first principal component that corresponds to the highest variation in the dataset, with major shifts after 26 and 48 h.

decrease in signal of extracts A and B, after comparable time intervals following the start of the analysis of each extract. Principal component analysis of the dataset shows that the LC–MS performance continuously changes, with a clear shift after 26 and 48 h (Fig. 8.4 plots samples from extract A; extract B showed an identical pattern). Overall, the majority of detectable metabolites did not significantly vary over time, with only slight increases in the average relative standard variation over time (11.18%, 12.77% and 14.37% after 24, 48 and 72 h, respectively; see Electronic Supplementary Material Figure S1 for trends of individual metabolites in extract A). However, a subset of 50 out of 235 metabolites was the source of the variability seen in the PCA score plot. These compounds showed either a signal intensity close to the minimum detection level and were thus not detected continuously or they were sensitive to degradation during storage in the autosampler. The latter can be identified as a consistently decreasing trend across time. Figure 8.5 clearly shows the oxidation of glutathione and formation of its oxidised form, glutathione disulfide, during storage in the autosampler, for both extracts A and extract B (note the 12-h time shift as expected). Trypanothione (Fig. 8.6) and its oxidised product, trypanothione disulfide, show an even faster oxidation rate, making quantification

of this marker metabolite impossible without special precautions. Other oxidation-sensitive products like cysteine, glutamylcysteine and glutathionylspermidine also show oxidation during autosampler storage.

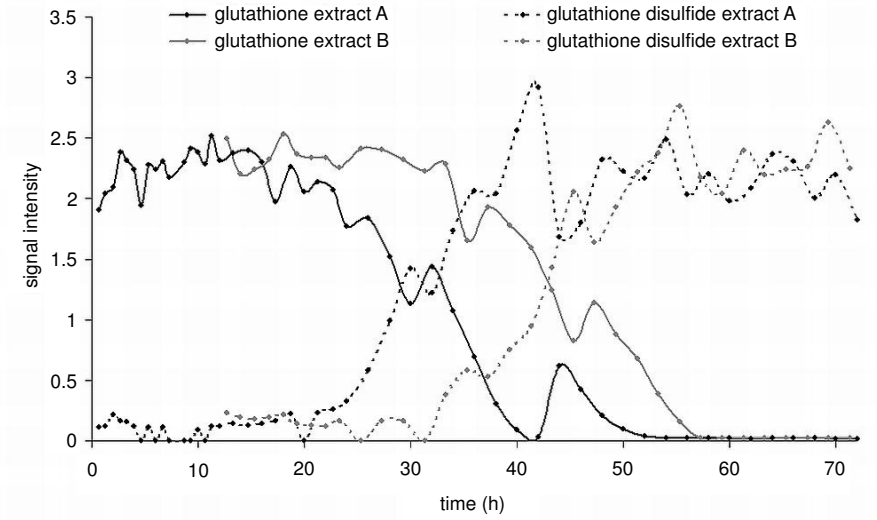


Figure 8.5: Levels of metabolites sensitive to oxidation during storage in the autosampler (4 °C), for extract A and extract B (measurement started 12 h after extract A). Intensity values (in counts) are scaled to unit variance. The signal of glutathione immediately drops due to formation of its oxidation product, glutathione disulfide, showing the necessity of oxygen deprivation during autosampler storage

These results clearly demonstrate the necessity for oxidation prevention. Nitrogen and argon gas were shown to prevent extracts from oxidation and degradation caused by atmospheric air (Fiehn, 2006; Lisec *et al.*, 2006). A deoxygenating step using nitrogen gas was consequently included in the sample preparation protocol. Deoxygenated samples show stable profiles of oxidation-sensitive metabolites (see Electronic Supplementary Material Figure S2).

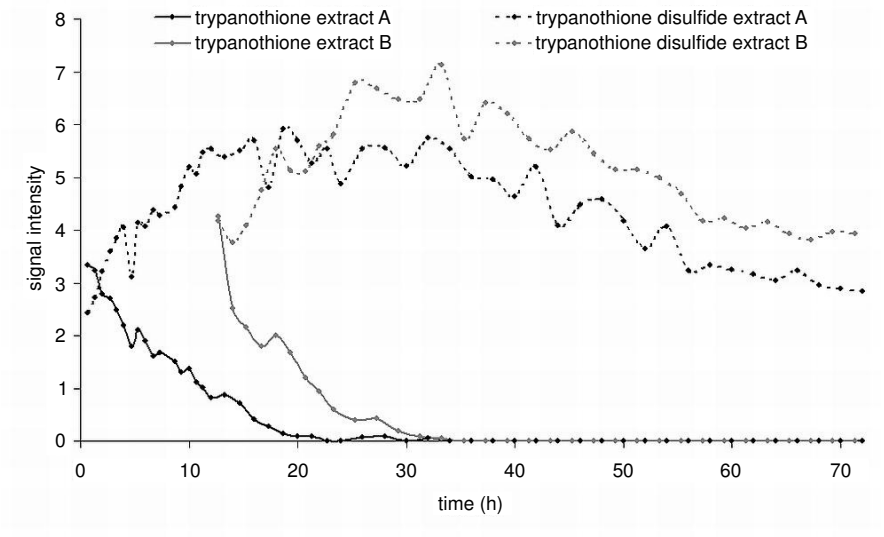


Figure 8.6: Levels of metabolites sensitive to oxidation during storage in the autosampler (4 °C), for extract A and extract B (measurement started 12 h after extract A). Intensity values (in counts) are scaled to unit variance. The signal of trypanothione immediately drops due to formation of its oxidation product.

8.2.5 Coverage of HILIC–Orbitrap analysis

To illustrate the final performance of our optimised protocol (Fig. 8.7), we compared the global metabolome of three different *L. donovani* strains. HILIC–Orbitrap analysis was performed in both negative and positive ionisation mode to maximise the coverage. All cultures were growth-synchronised to ensure harvest at the same growth stage.

Principal component analysis shows huge variation between the different strains on metabolome level, clearly demonstrating the potential of metabolomics in protozoa-related research, as described before (Scheltema *et al.*, 2010) (see Electronic Supplementary Material Figure S3). About 20–50% of the detected metabolome variation can be linked to differences between strains.

To assess the metabolite coverage obtained by our protocol, we generated a conservative list of detected metabolites using the following inclusion criteria: (a) reproducible signal intensity (RSD%<0.35 across the biological replicates), (b) consistent detection (detected in all samples; n=11), (c) peak shape quality (Coda-DW value>0.8), (d) retention time matching in positive and negative ESI mode, and (e) high mass accuracy (<1 ppm). We matched the resulting metabolite list to the

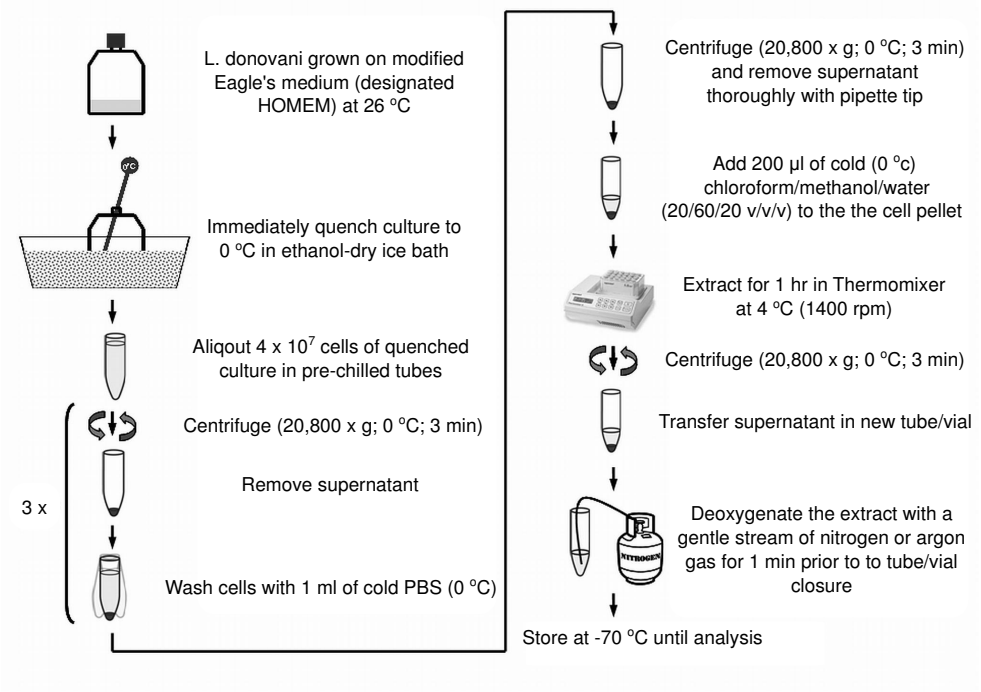


Figure 8.7: Final optimised protocol for the study of the intracellular metabolome of *Leishmania* parasites.

LeishCyc database (Doyle *et al.*, 2009), a biochemical pathway database for the closely related *Leishmania major*, and estimated the approximate coverage of the *L. donovani* metabolome achieved with our approach.

The LeishCyc database is still being updated and contains 566 metabolites so far; the entire *Leishmania* metabolome is predicted to contain about 1,101 metabolites (Chavali *et al.*, 2008). However, a cell-based metabolomics analysis encompasses the different compartments within a cell. If one corrects for the metabolites occurring in multiple cellular compartments, there are about 600 unique metabolites (Scheltema *et al.*, 2010). A total of 118 metabolites of the LeishCyc database were present in our list of putatively identified metabolites detected with our developed LC–MS approach, corresponding to 20% coverage of the predicted *Leishmania* metabolome. All identified LeishCyc metabolites are listed in the Electronic Supplementary Material Figure S4. As the LeishCyc database is not complete (e.g., phospholipids are not included), we also matched our list to more extensive databases like KEGG (Ogata *et al.*, 1999), LipidMAPS (Fahy *et al.*, 2007), Human Metabolome Database

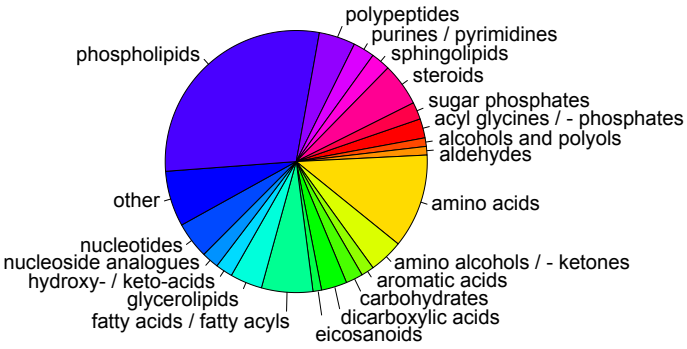


Figure 8.8: Pie chart representing coverage of metabolite classes using our optimised analytical protocol. Metabolite classes are gathered from the Human Metabolome Database (Wishart *et al.*, 2007). The class defined as other contains all metabolite classes that have only one or two representative compounds (i.e. pyridoxals and derivatives, quaternary amines, pterins, prenol lipids, indoles and derivatives) or metabolites that are not classified (e.g. ovothiol and trypanothione). A total of 379 metabolites were detected and putatively identified

(Wishart *et al.*, 2007) and a peptide database. Overall, a total of 379 mass signals (combining positive and negative ESI) are annotated using all databases combined. An overview of the metabolite classes represented in our data is shown in Fig. 8.8. The largest groups of annotated metabolites consisted of phospholipids (29%) and amino acids (12%). Also, a high number of fatty acids/fatty acyls (6%), steroids and steroid derivatives (5%), polypeptides (5%) and nucleotides (5%) could be annotated.

8.3 Conclusions

In this study, we aimed to develop a specific method for the analysis of the metabolome of *Leishmania* species. We have chosen to work with the promastigote life stage of *Leishmania* as this form is relatively easy to culture *in vitro* and has an extracellular lifestyle. Biologically speaking, it would also be interesting to study the clinical relevant life stage, amastigotes, but their intracellular lifestyle in mammalian macrophages hampers a reliable and accurate parasite-specific metabolite extraction and subsequent detection. Our final sampling protocol starts with a rapid quenching of promastigotes to 0 °C in their culture medium. A triplicate washing step with phosphate-buffered saline is required to remove the rich culture medium. We could demonstrate that cell leakage is absent during this extensive washing procedure. Cell

disruption and comprehensive metabolite extraction were found to be optimal using 4×10^7 parasites in 200 μl cold chloroform/methanol/water 20:60:20 (*v/v/v*) for 1 h at 4 °C. We also demonstrated that selected key metabolites show rapid oxidation during storage in the autosampler, which we try to minimise by including a deoxygenation step of the extract.

The obtained coverage of 20% of the predicted metabolome, involving metabolites from many different pathways, is a good starting point for the untargeted detection of metabolome changes in our large collection of *L. donovani* clinical isolates. However, a coverage of 20% compares unfavourably with what can be achieved by proteomic or transcriptomic analyses (Rosenzweig *et al.*, 2008; Saxena *et al.*, 2007; Scheltema *et al.*, 2010). But even if not all metabolites can be detected (due to low abundance, instability or incompatibility with the separation conditions), the broad coverage of the metabolome ensures that "marker metabolites" of most metabolic processes are included in our dataset. This will provide a unique insight into the cellular function of genetically and phenotypically diverse strains of this complex pathogen. The improvement of metabolite coverage using multiple analytical platforms is a major focus of our ongoing research efforts.

Acknowledgements

This work was supported by the GeMInI initiative of the Institute of Tropical Medicine. Ruben t'Kindt was supported by a Research Foundation Flanders Grant for a long stay in Strathclyde. Rainer Breitling was supported by an NWO-Vidi fellowship.

We are grateful to Prof. Malcolm McConville and his team at the Bio21 Institute at the University of Melbourne for their advice and guidance regarding the Leishmania metabolome extraction protocol.

Electronic supplementary material

Additional information as noted in text. This material is available free of charge via the Internet at `http://link.springer.com/content/esm/art:10.1007/s00216-010-4139-0/MediaObjects/216_2010_4139_MOESM1_ESM.pdf`.

Chapter 9

Metabolomic analysis of a synthetic metabolic switch in *Streptomyces coelicolor* A3(2)

Andris Jankevics[1,2,*], M. Elena Merlo[1,3,*], Marcel de Vries[4], Roel J. Vonk[4], Eriko Takano[3], Rainer Breitling[1,2]

Proteomics 11(24):4622-4631, 2011

1 Groningen Bioinformatics Centre, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Groningen, The Netherlands
2 Institute of Molecular, Cell and Systems Biology, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, United Kingdom
3 Microbial Physiology, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Groningen, The Netherlands
4 Centre for Medical Biomics, University Medical Centre Groningen, Groningen, The Netherlands
* Equal contribution

abstract

The global analysis of metabolism by liquid chromatography coupled to mass spectrometry is often hampered by a large amount of biological and technical variability. Here, we introduce an experimental and analytical strategy that can produce robust metabolome profiles in the face of this challenge. By applying a new computational approach based on concordance analysis to an extremely large number of analytical replicates, we are able to show that the overexpression of an antisense non-coding RNA targeting glutamine synthetase I results in a major reorganization of the metabolism of Streptomyces coelicolor, the model species of antibiotic-producing bacteria. We identified 97 metabolites with statistically significant reproducible dynamic behavior across the time series. The observed metabolic changes are very rapid, specific and widespread across metabolism, but focus on the nitrogen assimilation pathways. Our results demonstrate the power of highly replicated experimental designs for the robust characterization of metabolite dynamics. The identified global rearrangement of metabolism suggests the usefulness of RNA interference as an efficient strategy to manipulate the physiology of bacteria with wider biotechnological applicability in microorganisms.

Actinomycete bacteria are widely exploited in the biotechnological industry for their ability to produce a large variety of bioactive compounds of wide use in agriculture and medicine (Clardy *et al.*, 2006). The model species of the group, *Streptomyces coelicolor*, is known to produce a large number of secondary metabolites, including five well-characterized antibiotics (Bentley *et al.*, 2002; Challis and Hopwood, 2003; Fiehn, 2002). The production of secondary metabolites is generally associated with a global metabolic switch, often induced by entering the stationary phase of growth upon nutrient limitation (Alam *et al.*, 2010; Horinouchi, 2007; Nieselt *et al.*, 2010; Ruiz *et al.*, 2010). However, this switch is not well understood and characterized. A method to induce a metabolic switch without nutrient limitation would be of great biological and biotechnological interest.

Here, we explore the ability of an overexpressed non-coding ribonucleic acid (RNA) to induce a "synthetic" metabolic switch. The regulatory function of bacterial small non-coding RNAs has recently been subject of several studies (Hebert *et al.*, 2008; Lioliou *et al.*, 2010; Thomason and Storz, 2010; Waters and Storz, 2009), and they have also been identified in *S. coelicolor* (D'Alia *et al.*, 2010).

The glutamine synthetase, GSI, plays a major role in nitrogen assimilation in *S. coelicolo*r by catalyzing the conversion of ammonia and glutamate to form glutamine

(Amon *et al.*, 2010; Reuther and Wohlleben, 2007). An intragenic chromosomal cis-encoded non-coding RNA in the gene glutamine synthetase I gene (*glnA*), which encodes for GSI, was predicted by computational analysis and its physiological role explored in *S. coelicolor* by D'Alia *et al.* (2010). Overexpression of the non-coding RNA resulted in a decrease in growth, protein synthesis and antibiotic production.

To determine whether the overexpression resulted in a true metabolic switch, i.e. globally affected the metabolic status of the cells, we applied large-scale untargeted metabolomics. For this analysis, the development of new statistical approaches and the collection of an unusual amount of data on biological, analytical and technical replicates were essential to distinguish real effects due to the antisense overexpression from spurious fluctuations in cellular metabolism. The strategies developed here will be generally useful for microbial metabolomics.

9.1 Material and methods

9.1.1 Bacterial strain, media and growth conditions

Experiments were performed using *S. coelicolor* M145/pTE313, which contains a thiostrepton-inducible non-coding RNA expression construct targeting glnA described previously (D'Alia *et al.*, 2010).

Bacteria were grown in 50 ml liquid medium (Nieselt *et al.*, 2010) in siliconized 250 ml flasks (siliconizing agent (silanization solution II) for the cultivation flasks was from Fluka, Sigma-Aldrich, Netherlands) containing stainless-steel springs. Cultures were grown in a shaking incubator at 220 rpm at 30 °C.

Expression of the non-coding RNA was induced when the cultures reached an OD_{450} of 0.5 by addition of 12 μL of thiostrepton (Merck KgaA, Germany) dissolved in DMSO (Boom, Netherlands) at the final concentration of 5 μg/mL.

As a negative control, we used the strain M145/pTE313 with addition of 12 μL of DMSO to compare the same bacterial strain therefore avoiding unwanted changes in physiology that can be seen between different *Streptomyces* strains. Moreover, the inducing chemical, thiostrepton, has a very limited number of molecular targets in the *Streptomyces* genome (Ali *et al.*, 2002; Murakami *et al.*, 1989).

9.1.2 Metabolome sampling

Metabolome samples were collected at 0, 2, 4, 6 and 8 h after induction. Cells from 25 ml culture were collected on a 0.45-μm filter by vacuum filtration and washed twice with 25 ml of 2.63% NaCl solution. For cell quenching, the filter with the collected

cells was quickly moved into 60% methanol (HPLC-grade, Boom, Netherlands) solution pre-chilled at –20 °C and the solution was frozen in liquid nitrogen. The transfer time of filtrated cell material to the quenching solution was less than 10 s. Samples were stored at –80 °C until metabolite extraction was performed.

9.1.3 Metabolite extraction

Metabolites were extracted by three freeze–thaw cycles. Cells from –80 °C were thawed in ethanol bath at –20 °C (∼15 min), vortexed vigorously for 1 min and, right afterwards, frozen in liquid nitrogen for 5 min. The cycle was repeated three times. After the third cycle, the samples were centrifuged at 4500 rpm for 10 min at –9 °C. The supernatant (cell extract) was collected and stored at –80 °C until LC–MS analysis.

9.1.4 LC–Orbitrap MS analysis

The cell extracts were analyzed by liquid chromatography coupled to a high-resolution LTQ Orbitrap XL mass spectrometer (Thermo Fisher Scientic, Germany).

Two chromatographic columns were used: a reversed-phase Shim-pack XR-ODS C18 column (Achrom, Belgium) (3.0 × 75 mm, 2.2 μm, Shimadzu) and a ZIC-HILIC column (Achrom, Belgium) (150 × 2.1 mm, 3.5 μm, Merck Sequant AB) fitted with a ZIC-HILIC PEEK guard column (Achrom, Belgium) (15 × 1.0 mm; 5 μm, Merck Sequant AB). The sample volume injected was 5 μl for both columns.

For the C18 column, the flow rate was set to 0.6 ml/min; the mobile phase consisted of (A) 0.1% formic acid in water and (B) 0.1% formic acid in acetonitrile. A gradient of 18 min was used. The elution of solvent B started at 2% for the first 2 min and was increased to 95% within 8 min. This composition was maintained for 2 min, after which the elution of B was decreased to 2% within 1 min. To re-equilibrate the system, the elution of B was held at 2% for 5 min.

For the ZIC-HILIC column, the flow rate was set to 0.1 ml/min; as buffers, (A) 0.1% formic acid in ACN and (B) 0.1% formic acid in water were used. A gradient of 40 min was applied. The elution of solvent A was set to 80% as starting condition. The elution of solvent B was increased to 40% within 6 min and maintained at 40% for 12 min, after which solvent B was increased to 90% in a 4-min interval. This composition was held for 2 min after which B was decreased to 20% in 2.5 min. The gradient was held at 20% B for 13.5 min to re-equilibrate the system.

The system was operated with electrospray ionization source in positive and negative mode. Full-scan spectra were obtained over an m/z range of 50–1000 Da.

ULC-grade ACN, formic acid and water were purchased at Biosolve (Netherlands).

9.1.5 Data processing

Raw data files from the mass spectrometer were converted into the mzXML format by the ReAdW.exe utility (a tool of the Trans-Proteomic Pipeline software collection, downloaded from `http://tools.proteomecenter.org/wiki/index.php?title=Software:ReAdW`).

The CentWave (Tautenhahn *et al.*, 2008) feature detection algorithm from the XCMS (Smith *et al.*, 2006) package was used on each individual data file. Further processing was handled by the flexible data processing pipeline mzMatch (Scheltema *et al.*, 2011) (`http://mzmatch.sourceforge.net`), performing noise removal (Windig, 2004) and several steps of signal filtering and peak matching. The first matching step involved aligning of the chromatographic features between technical replicates of a single sample. Peaks that were not detected in all technical replicates were discarded from further analysis. In the second matching step, the chromatographic peaks, which were combined in single files containing technical replicates in the previous matching step, were aligned to each other for all five time points and across sample groups (thiostrepton-induced and negative control). After combining several measurements in a single file, there were still peak sets that do not include peaks from every sample. Such gaps were filled by extracting ion chromatograms within the retention time and mass window of the given peak set directly from the raw data files. The same processing steps (matching of peak sets and gap filling) were then repeated to combine multiple biological replicates.

Derivative signals (isotopes, adducts, dimers and fragments) were automatically annotated by correlation analysis on both signal shape and intensity pattern, as described by Scheltema *et al.* (2009). The derivative signals were removed before further statistical analysis, as they would give excessive weight to differentially expressed analytes with many mass derivatives.

Putative identifications were made by matching the detected masses to a number of online databases: ScoCyc (Bentley *et al.*, 2002), MetaCyc (Caspi *et al.*, 2008), KEGG (Kanehisa *et al.*, 2002), LipidMAPS (Fahy *et al.*, 2007), Human Metabolome Database (Wishart *et al.*, 2007), a contaminant database (Keller *et al.*, 2008) and a custom-made general peptide database containing all di- and tri-peptides of the proteinogenic amino acids. The identifications for amino acids were confirmed by comparison to MS data obtained from a standard mixture (A9906, Sigma-Aldrich) containing amino acids and related compounds.

Kendall's non-parametric W statistics (Kendall and Smith, 1939; Legendre, 2005)

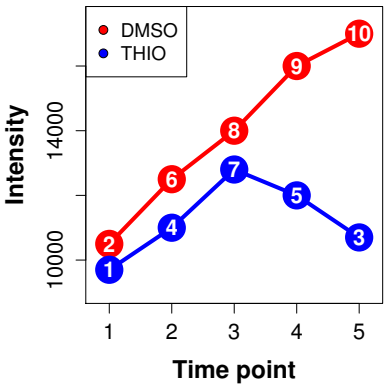


Figure 9.1: Scheme of assigning ranks to the observed intensity values in one biological replicate (red: negative control samples, blue: thiostrepton-induced samples).

was chosen as an effective measure to choose metabolites, which are sufficiently reproducible in their dynamic behavior between biological replicates. First, the observations for each biological replicate (between the control and the induced samples) were ranked according to their signal intensity (see Fig. 9.1 for an example of the ranking procedure). These ranks were then used to calculate Kendall's W across all biological replicates as follows:

$$W = \frac{12 \sum_{i=1}^{n} (\sum_{j=1}^{m} r(x_{i,j}) - \frac{1}{2} m(n+1))^2}{m^2(n^3 - n)} \tag{9.1}$$

where W is the Kendall's coefficient of concordance, n is the number of observations (number of time points $\times$ number of technical replicates), m is the number of biological replicates, $x_{i,j}$ is the intensity in observation i of the j-th biological replicate, $r(x_{i,j})$ is the rank of $_{xi,j}$ in the vector x_j of intensities in j-th biological replicate.

The corresponding p-value, i.e. the probability of observing such a high W in random data, is also calculated. Multiple testing correction was applied to the p-values from Kendall's concordance test according to the Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995).

This method focuses on compounds with dynamically changing intensities and will not report metabolites that are constant across the time course; the latter have to be identified by independent filtering criteria. We calculated a stability score (Sc) for each metabolite as follows:

$$Sc = \frac{1}{t \cdot m} \sum_{j=1}^{m} \frac{\sum_{a=1}^{t} (D_{j,a} - T_{j,a})^2}{(\frac{1}{2 \cdot t} \sum_{a=1}^{t} (D_{j,a} + T_{j,a}))^2} \tag{9.2}$$

where Sc is the Stability score, a smaller value indicates a better agreement between time courses, m is the number of biological replicates, t is the number of time points in growth curve, $D_{j,a}$ is the average intensity value of the j-th biological replicate at the a-th time point from negative control samples, $T_{j,a}$ is the average intensity value of the j-th biological replicate at the a-th time point from thiostrepton-induced samples.

In the equation for the stability score, the numerator favors compounds that show small differences between control and induced samples, whereas the denominator favors compounds with minimal variation across the time course.

Identified metabolites were mapped onto metabolic pathways chart from the KEGG repository (Kanehisa *et al.*, 2002) using the Interactive Pathways Explorer tool (Letunic *et al.*, 2008) followed by manual curation to match compounds identified from the ScoCyc database (Bentley *et al.*, 2002).

All statistical analyses and graphical routines were handled in R (R Core Team, 2012). The R code for the reading and writing of data from/to the PeakML file format, the XML representation of processed data produced by the mzMatch pipeline, is available from the authors upon request.

		Technical replicates			
		ZIC-HILIC MS		**C18 MS**	
	Growth curve	**Positive ESI**	**Negative ESI**	**Positive ESI**	**Negative ESI**
	1	3			
Biological	2	3			
replicates	3	3	3	3	3
(5 time	4	3	3	3	3
points each)	5	3	3	3	3
	6	6	4	4	4

Table 9.1: List of analytical, technical and biological replicates analyzed. The different chromatography and mass spectrometry setups are described in "Material and methods" section.

9.2 Results and discussion

To characterize the metabolic consequences of overexpressing an antisense non-coding RNA targeting *glnA*, we obtained a large, highly replicated set of metabolomic profiles (Table 9.1).

Samples were collected from six independent growth curves and analyzed in at least three technical replicates on two different chromatographic columns (C18, zicHILIC) and using two ionization modes (positive and negative). Each growth curve consisted of 5 time points. This means that for each compound, up to 300 data points are available in the control and treated condition (five time points × two ionization modes × two columns × up to six growth curves × at least three technical replicates). This exceptional level of replication, together with a very rigorous filtering algorithm (as described in "Material and methods" section), allowed us to identify reproducible metabolic trends despite the well-known biological variability of *Streptomyces* batch cultures (Betts and Baganz, 2006; Büchs, 2001). Figure 9.2 shows the effects of each subsequent filtering step. The goal of our study was to focus only on metabolites that are sufficiently reproducible in their behavior between biological replicates.

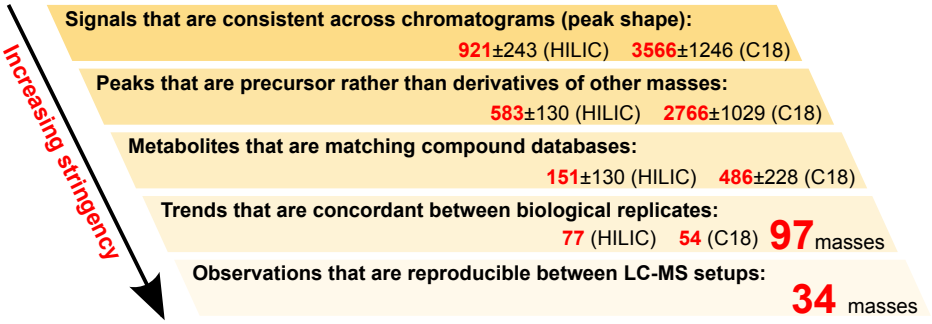


Figure 9.2: A hierarchy of filters showing efficiency of carefully selected data processing pipeline, which allows to remove irreproducible signals from the data set.

Even subtle differences are reliably identified in this filtered set of metabolites, as can be seen in the example in Fig. 9.3a, which shows the time course for a mass of 118.02661 Da putatively identified as succinate. This mass was detected only in negative mode but on the two different chromatographic columns. Even though the absolute intensities were low and the measurements were done months apart using different instrumental conditions and considering the large variation between

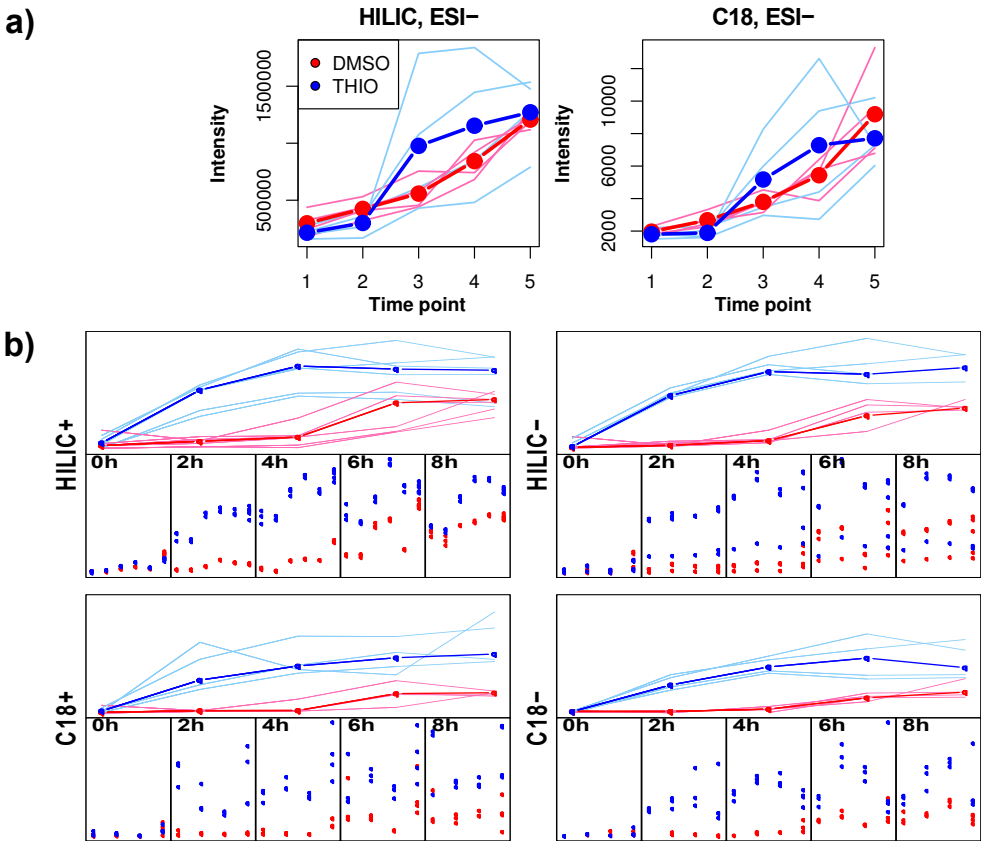


Figure 9.3: (**a**)Succinate (measured mass 118.02661 Da) as an example of excellent reproducibility between biological replicates in the filtered data set. (**b**) Representation of the technical, biological and analytical variability between the replicates for a mass of 146.06913 Da, putatively identified as glutamine. This compound was detected in all four analytical conditions (HILIC+, HILIC–, C18+ and C18–) in the five measured time points. Each subplot (tile) represents the observations for one time point in multiple biological replicates (six for HILIC+ data and four for the rest), where red and blue dots correspond to the negative control and thiostrepton-induced samples, respectively. The thin lines in the summary plots at the top represent the average for each biological replicate, and the bold line the median across all replicates. Variation in the detected signal among biological replicates and within some analytical conditions is very high, but nonetheless the details of the dynamic trends are identical in all four analytical methods, even though measurements were obtained several weeks apart. The concordance analysis described in the text was developed as a principled method to identify compounds with such reproducible behavior.

replicates (shown as light lines in the figure), the small differences in the detected trends are exactly the same.

This excellent level of reproducibility represents a major achievement in metabolomics, which not only suffers from technical noise introduced by machine drifts in sensitivity and resolution, but also from a large amount of biological variability. Figure 9.3b shows the reproducible trends identified after data processing and filtering for a mass with highly concordant behavior. This high reproducibility is achieved despite the clearly visible intensity variation between technical and, in particular, biological replicates.

Before further data processing all peaks with corrected p-values larger than 0.10 were excluded (if p-values were higher than 0.1 in one analytical condition, but lower in another, they were retained in the data set). This procedure left us with 97 putatively identified unique compounds that were concordant across biological replicates (77 metabolites for HILIC samples, and 54 for the C18 column); 34 of them were detected consistently in at least two different analytical conditions (see Supporting Information 1 for mass chromatograms, putative identifications, Kendall's W statistics, and trends of all selected individual metabolites). The list of the 34 most reproducible compounds detected in at least two different analytical conditions is shown in Table 9.2.

Surprisingly, we did not detect any metabolites that showed a constant behavior during the time course either in the negative control or in thiostrepton-induced samples. The majority of the peaks identified with stable intensity trends across the time series (stability score calculated as described in "Material and methods" section.) are contaminants typical for LC–MS analysis (Keller *et al.*, 2008) or spuriously detected compounds with intensities close to the detection limit. This stability of signals for ubiquitous contaminants confirms the reproducibility of the relative quantification. A list of the detected peaks with mass chromatograms, putative identifications, intensity trends and stability score values is available in Supporting Information 2.

To obtain a general picture of the metabolite dynamics after induction of the synthetic metabolic switch by expression of the antisense RNA, the 97 most stringently filtered putatively identified compounds were visualized by hierarchical clustering (if a metabolite was detected in more than one analytical condition, then the peak with the smallest p-value was selected), using Kendall's correlation coefficient of the signal intensities as the similarity metric. Only a limited number of distinct dynamic patterns were detected (Fig. 9.4a).

Each of the 97 putatively identified metabolites responded to the synthetic switch with a distinctive trend, and three major dynamic behaviors were recognized when

Measured mass	Predicted chemical formula	Mass error, ppm	Analytical conditions[a] 1	2	3	4	Putative identification[b]
146.06913	$C_5H_{10}N_2O_3$	0.1	<0.01	0.01	0.05	0.08	Glutamine
189.06373	$C_7H_{11}NO_5$	–0.1	<0.01	0.01	0.05	0.08	*N*-Acetyl-glutamate
275.11162	$C_{10}H_{17}N_3O_6$	0.4	<0.01	0.01	0.05	0.09	Glu-Gln
246.12153	$C_{10}H_{18}N_2O_5$	0.2	<0.01	0.01	0.05	0.08	Leu-Asp
142.07424	$C_6H_{10}N_2O_2$	–0.1	<0.01	0.01	0.05	0.08	Ectoine
158.06912	$C_6H_{10}N_2O_3$	0.1	<0.01	0.02	0.05	0.08	Hydroxyectoine
89.04771	$C_3H_7NO_2$	–0.3	<0.01	0.03	>0.1	>0.1	Alanine
260.13716	$C_{11}H_{20}N_2O_5$	0.2	<0.01	0.01	>0.1	0.08	Glu-Leu
160.08473	$C_6H_{12}N_2O_3$	0.4	0.04	0.02	>0.1	>0.1	Ala-Ala
347.06322	$C_{10}H_{14}N_5O_7P$	–0.4		0.01	0.05	0.08	AMP
190.09523	$C_7H_{14}N_2O_4$	0.7	<0.01	0.01	0.08		Diaminopimelate
175.09551	$C_6H_{13}N_3O_3$	1	<0.01	0.01	0.07		Citrulline
199.02466	$C_4H_{10}NO_6P$	–0.5		0.01	0.07	0.08	*O*-Phospo-homoserine
172.08468	$C_7H_{12}N_2O_3$	0.6	<0.01	0.01	0.1		Gly-Pro
232.10582	$C_9H_{16}N_2O_5$	0.4	<0.01	0.01	0.07		*N*-Succinyl-ornithine
161.06868	$C_6H_{11}NO_4$	0.8	<0.01	0.01	0.11		*O*-Acetyl-homoserine
342.1166	$C_{12}H_{22}O_{11}$	–1.1		0.01		0.08	Sucrose, threhalose
115.06332	$C_5H_9NO_2$	0.1	<0.01		0.05		Proline
181.07379	$C_9H_{11}NO_3$	0.6	<0.01		0.05		Tyrosine
118.02661	$C_4H_6O_4$	0		0.01		0.08	Succinate
324.03616	$C_9H_{13}N_2O_9P$	–0.9		0.02		0.08	UMP
165.0789	$C_9H_{11}NO_2$	0.5	<0.01		0.05		Phenylalanine
105.04256	$C_3H_7NO_3$	0.3	<0.01	0.01			Serine
117.07896	$C_5H_{11}NO_2$	0.2	<0.01		>0.1		Valine
290.04062	$C_7H_{15}O_{10}P$	–1.2		0.01		0.08	S1P
131.09464	$C_6H_{13}NO_2$	–0.1	<0.01		0.06		Leucine
229.08834	$C_9H_{15}N_3O_2S$	0.7	<0.01	0.02			Ergothioneine
314.05188	$C_8H_{15}N_2O_9P$	–1.1		0.01		0.09	N-Formyl-GAR
380.09023	$C_{21}H_{16}O_7$	–1.7		0.01		0.08	Tcm-F1-OMe
133.03749	$C_4H_7NO_4$	0.2	<0.01		>0.1		Aspartate
174.10035	$C_7H_{14}N_2O_3$	0.5	<0.01	0.02			*N*-Acetyl-ornithine
134.0216	$C_4H_6O_5$	–0.6		0.02		>0.1	Malate
119.05821	$C_4H_9NO_3$	0.3	<0.01	0.04			Threonine
175.04798	$C_6H_9NO_5$	0.5	<0.01	0.04			*N*-Acetyl-aspartate

Table 9.2: List of the most significant concordant metabolites between biological replicates. Complete details for every metabolite, including chromatograms, are shown in Supporting Information 1. a) 1 – HILIC, ESI+; 2 – HILIC, ESI–; 3 – C18, ESI+; 4 – C18, ESI– (the multiple testing-corrected *p*-value of the concordance test is shown for each condition). b) Identifications are only putative guides for initial exploration, based only on exact mass information. Alternative isomers should also be considered for the interpretation. S1P – Sedoheptulose 1-phosphate, N-Formyl-GAR – 5'-Phosphoribosyl-*N*-formylglycinamide, Tcm-F1-OMe – Tetracenomycin F1 methylester.

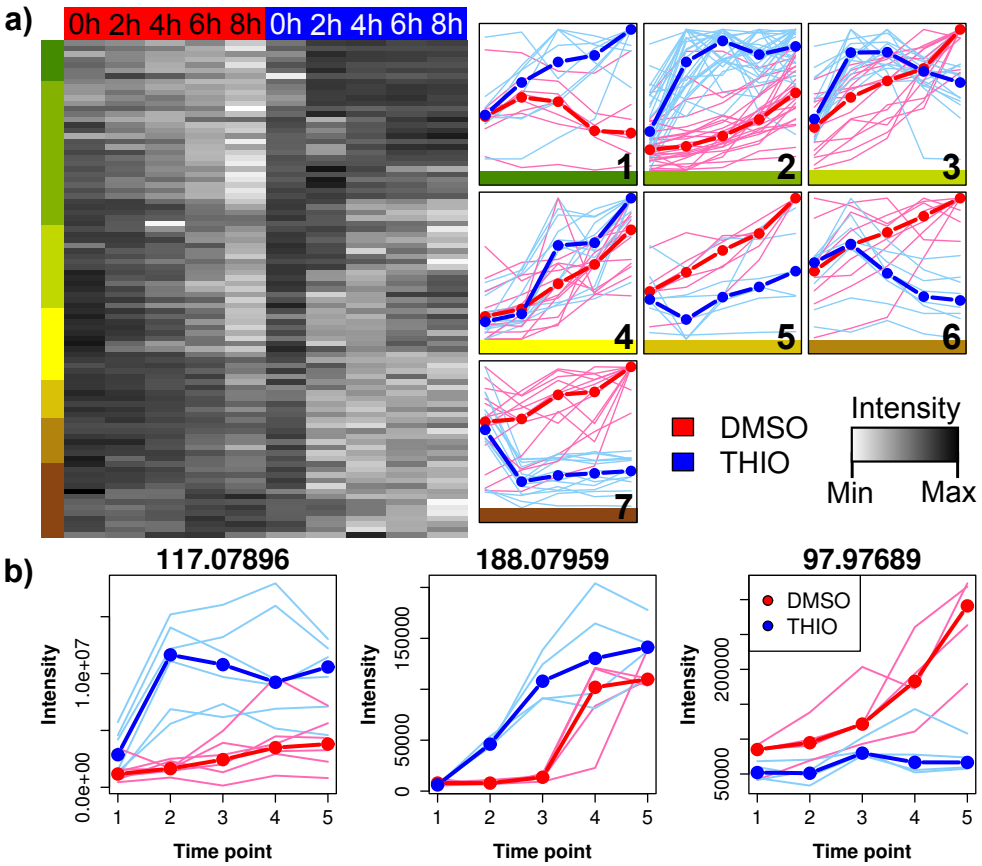


Figure 9.4: (**a**) Heatmap (left panel) and intensity trends (right panels) for the 97 most reproducible dynamically changing compounds. The metabolites were subdivided into seven clusters after the hierarchical clustering, as indicated by the color bars on the left, and the dynamics of the individual genes, as well as the average of each cluster, are shown on the right (intensities scaled to a maximum value of 1 for each metabolite). (**b**) Examples of dynamic trends of three putatively identified compounds. On the left, a mass of 117 Da (putatively identified as valine) is accumulated immediately after the switch in the thiostrepton-induced samples. In the center, a mass of 188 Da (acetyl-glutamine) shows a more delayed response to the non-coding RNA induction. On the right, the intensity levels for a mass of 97 Da (phosphoric acid) changes more gradually along the time series.

comparing the non-coding RNA overexpressing samples to the uninduced control (Fig. 9.4)a. The first group of metabolites (clusters 1, 2 and 3) shows a specific accumulation in the thiostrepton-induced samples; the second group (clusters 5, 6

and 7) is specifically depleted upon induction of the non-coding RNA expression; the third, smallest group (cluster 4) shows a dynamic change over time, but has very similar levels in control and induced samples.

Among the members of each group, the timing of the response to the non-coding RNA expression varies strikingly. For some metabolites, the abundance changes very rapidly after induction of non-coding RNA expression (e.g. mass 117.0789 Da in Fig. 9.4b). In others, the metabolic response was delayed (e.g. mass 188.07959 Da in Fig. 9.4b, or it occurred gradually along the time series (e.g. mass 97.97689 Da in Fig. 9.4b).

Of the 97 highly concordant metabolites, 35 could be assigned to metabolic pathways. In Fig. 9.5, the dynamic behavior of 35 putatively identified metabolites is shown projected onto a global map of the metabolic network of *S. coelicolor*. The left panel shows the behavior in untreated samples: The majority of the compounds are slowly accumulated over time; only one metabolite was depleted and a few were detected at a constant level during growth. In the treated cultures (right panel), metabolite dynamics are far more varied, most importantly showing rapid transient responses for the majority of the metabolites.

When comparing the trends detected in the treated samples to the control strain (Fig. 9.5), it is noteworthy that the strongest metabolic response to the overexpression of the non-coding RNA against GSI is centered on the nitrogen assimilation network, as expected. However, further strong changes are observed in many scattered areas of metabolism, indicating that the antisense overexpression has indeed resulted in a general metabolic switch associated to a wide perturbation of the metabolism of *S. coelicolor*.

The detected participants of the nitrogen assimilation network are very highly ranked in the concordance analysis (i.e. show robustly reproducible trends), but their response to the metabolic switch is diverse. Glutamate and glutamine, which are the substrates of GSI, are surprisingly both accumulating in the thiostrepton-induced samples but, while glutamine is rapidly increasing over time, glutamate shows only a transient accumulation within the first two time points. On the other hand, acetyl-glutamate, acetyl-ornithine and citrulline are depleted after induction of the metabolic switch. Carbamoyl-aspartate is instead accumulated in the first two hours after induction and then it is slowly depleted. The accumulation of glutamine is quite unexpected as the antisense overexpression results in a decrease of glutamine synthethase I levels (D'Alia *et al.*, 2010). *S. coelicolor* possesses five different glutamine synthethase genes (Reuther and Wohlleben, 2007) of which only two are directly involved in nitrogen assimilation: *glnA* (encoding for GSI, the target of the antisense

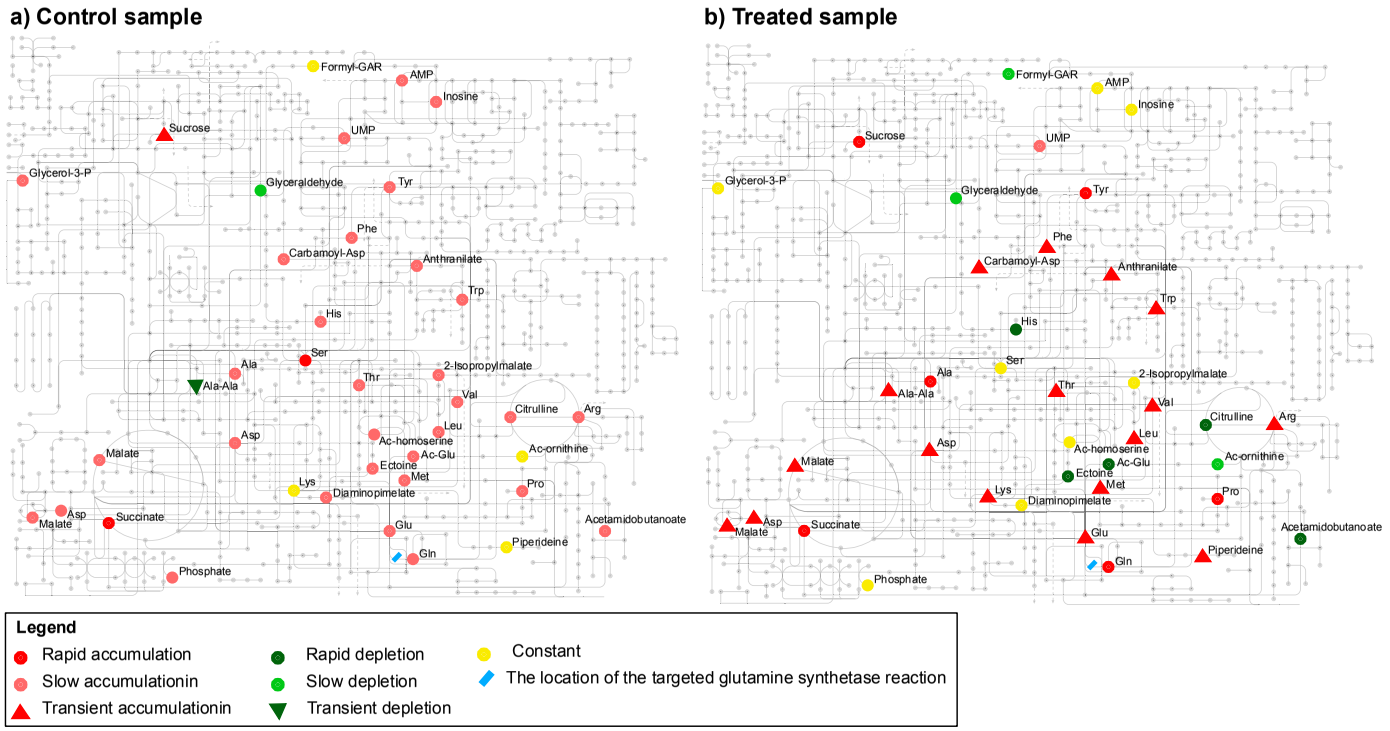


Figure 9.5: Metabolic pathways map of the highly concordant metabolites showing that the antisense overexpression has resulted in a general metabolic switch reflected in a widespread perturbation of the metabolome. The behavior of the putatively identified compounds is compared between the control samples (left panel) and the treated samples (right panel).

RNA) and glutamine synthetase II gene (*glnII*) (encoding for GSII). While GSI is considered the housekeeping glutamine synthetase in *S. coelicolor*, GSII was shown to be mainly active during morphological differentiation and development. However, GSI and GSII are able to complement each other in diverse culture conditions as was shown for *glnA* and *glnII* deletion mutants (Fink *et al.*, 1999). High expression for both GSI and GSII was earlier detected using the same medium as used here (Nieselt *et al.*, 2010), but at the moment we have no additional data indicating the possible complementary role of GSII after decreasing levels of GSI. On the other hand, the non-coding RNA may have off-target effects, or the regulatory network controlling nitrogen homeostasis may be so densely integrated that the effects of a local perturbation are not predictable based on the decreased activity of a single enzyme.

Amino acids and dipeptides are mostly transiently accumulated in the treated cells, with the exception of histidine and serine (Fig. 9.5). It is noteworthy that these dramatic changes, like those of the metabolites of the nitrogen assimilation network, are detected already within two hours after induction and before any growth delay is visible. This suggests that the changes in the metabolism of *S. coelicolor* are indeed directly and specifically caused by the non-coding RNA expression, and that they are not the unspecific result of a decrease in growth. Free phosphate, AMP, glycerol-3-phosphate, which were slowly accumulated in the control strain, are detected at constant intensity in the treated samples.

The fate of the osmoprotectants ectoine and hydroxyectoine closely follows the trend shown by their nitrogen-donor glutamate. In the control cultures, when glutamate increases over time, the two osmoprotectants are also accumulated. On the other hand, in the thiostrepton-induced samples, ectoine is strongly depleted after the antisense induction, and it seems to be totally converted to hydroxyectoine, which is transiently accumulated with a pattern similar to glutamate. In this case the level of hydroxyectoine accumulation may reflect the growth cessation caused by the induction of *glnA* non-coding RNA. The accumulation of hydroxyectoine and not ectoine was also observed in salt adaptation conditions (Kol *et al.*, 2010). There are several known links between ectoine biosynthesis and nitrogen metabolism, as bacteria can use ectoine as a nitrogen source after scavenging it from the external environment (Jebbar *et al.*, 2005). Ectoine is synthesized from aspartate and obtains an additional amino group from glutamate. Furthermore, the product of SCO1863, the first gene of the ectoine biosynthesis operon, is thought to be a nitrogen storage protein (Lewis *et al.*, 2011) and seems to be regulated by the global nitrogen metabolism regulator GlnR (Tiffert *et al.*, 2008). It is also important to realize that the medium used

in our study has a relatively high salt concentration (0.32 mM monosodium glutamate); this results in an increase in overall levels of osmoprotectants, making changes in ectoine levels easily detectable.

The species of the actinomycete genus *Streptomyces* are commonly known as antibiotic factories due to their capability of producing a large collection of secondary metabolites, including 70% of the antibiotics used in the clinical market (Hopwood, 2007; Weber *et al.*, 2003). The production of secondary metabolites starts in the secondary phase of metabolism when cells stop generating biomass and spend their energy to produce small chemical compounds as antibiotics and pigments. The best-known inducers causing cells to switch their metabolism from the primary to the secondary phase of growth are the limitation of nutrients or a drastic change in environmental conditions (e.g. pH variation) (Bode *et al.*, 2002; Scherlach and Hertweck, 2009).

Our results have shown that by the overexpression of specific antisense small RNAs we could generate a general metabolic switch well before nutrients were depleted. Obtaining such a rapid and wide perturbation of metabolism suggests that this post-transcriptional gene silencing technique can be an efficient way to manipulate cell physiology. This result was made possible by collecting an unusually large amount of replicated data, considering the simple basic question (i.e. which metabolites are affected by non-coding RNA overexpression?).

9.3 Concluding remarks

We have been able to manipulate the physiology of *S. coelicolor* by engineering an artificial metabolic switch using the overexpression of a specific non-coding RNA targeting GSI. Activating the engineered metabolic switch resulted in a major reorganization of metabolism, revealing the general utility of non-coding RNAs as a tool for the synthetic biology of actinomycetes (Medema *et al.*, 2011).

Our metabolome analysis was carried out on a unique data set with an unprecedentedly large amount of replication at multiple levels: biological replicates, technical ones, and analytical replicates using different technologies. This allowed us to apply a particularly stringent filtering of the data without losing informative signals. By applying concordance analysis we were able to detect reliable changes for 97 putatively identified compounds that are significantly concordant and reproducible across the large number of technical, analytical and biological replicates. This highly replicated experimental design will be of interest to other challenging metabolomics applications that are encumbered by a large amount of biological variation in metabolism.

Acknowledgements

The authors acknowledge Alexander Wentzel and Per Bruheim at SINTEF and NTNU (Trondheim, Norway) for providing the sample preparation protocol.
A.J. is supported by an NWO-Vidi award to R.B. M.E.M. is funded by a 4×4 Ubbo Emmius scholarship and E.T. by a Rosalind Franklin Fellowship, both from the University of Groningen. R.J.V. was supported by an investment grant from NWO.

The authors have declared no conflict of interest.

Supporting Information

Additional information as noted in text. This material is available free of charge via the Internet at
`http://onlinelibrary.wiley.com/doi/10.1002/pmic.201100254/suppinfo`.

Chapter 10

Stable isotope-assisted metabolomics for network-wide metabolic pathway elucidation

Darren J. Creek[1,2], Achuthanunni Chokkathukalam[3], Andris Jankevics[3,4], Karl E. V. Burgess[1], Rainer Breitling[3,4], Michael P. Barrett[1]

Anal. Chem. 84(20):8442-8447, 2012

1 Wellcome Trust Centre for Molecular Parasitology, Institute of Infection, Immunity and Inflammation, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, United Kingdom
2 Department of Biochemistry and Molecular Biology, Bio21 Molecular Science and Biotechnology Institute, University of Melbourne, Melbourne, Australia
3 Institute of Molecular, Cell and Systems Biology, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, United Kingdom
4 Groningen Bioinformatics Centre, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Groningen, The Netherlands

abstract

The combination of high-resolution LC–MS-based untargeted metabolomics with stable isotope tracing provides a global overview of the cellular fate of precursor metabolites. This methodology enables detection of putative metabolites from biological samples and simultaneous quantification of the pattern and extent of isotope labeling. Labeling of Trypanosoma brucei cell cultures with 50% uniformly ^{13}C-labeled glucose demonstrated incorporation of glucose-derived carbon into 187 of 588 putatively identified metabolites in diverse pathways including carbohydrate, nucleotide, lipid, and amino acid metabolism. Labeling patterns confirmed the metabolic pathways responsible for the biosynthesis of many detected metabolites, and labeling was detected in unexpected metabolites, including two higher sugar phosphates annotated as octulose phosphate and nonulose phosphate. This untargeted approach to stable isotope tracing facilitates the biochemical analysis of known pathways and yields rapid identification of previously unexplored areas of metabolism.

Recent advances in metabolomics technologies have enabled the analysis of metabolism on a global scale, allowing detection and semiquantification of hundreds of low molecular weight (<1000 Da) molecules in a single experiment (Dunn *et al.*, 2011b; Kueger *et al.*, 2012). However, interpretation of these results remains dependent on our knowledge of active metabolic pathways, which is far from complete for many organisms (Chen and Vitkup, 2007; Clasquin *et al.*, 2011; Dunn *et al.*, 2011b). The *ab initio* identification of novel pathways remains a challenge. Stable isotope tracing is an established technique for determining the fate of individual metabolites, but it generally requires targeted analysis of predicted products (Winder *et al.*, 2011; Zamboni and Sauer, 2009). The common technology that underlies both stable isotope tracing and metabolomics (i.e., mass spectrometry) provides the opportunity to combine these two approaches, allowing network-wide investigation of metabolic pathways.

High-resolution mass spectrometry offers great potential for untargeted metabolomics, as high mass accuracy often (but not always) allows direct formula assignment for each detected mass (m/z) (Breitling *et al.*, 2006b; Creek *et al.*, 2012b; Dunn *et al.*, 2011b; Kueger *et al.*, 2012). The complexity of liquid chromatography–mass spectrometry (LC–MS) data and the existence of many isobaric metabolites hinders unambiguous identification of all metabolite peaks, but recent advances in data processing allow automated identification and annotation of many metabolites with improved confidence (Brown *et al.*, 2011; Creek *et al.*, 2012a).

A remaining limitation regarding biological interpretation of metabolomics data is

the dependence on predetermined metabolic pathways, often overlooking the potential involvement of novel pathways. Even in well-defined organisms, detected metabolites may not derive from the classical pathways, as metabolites may be imported from an exogenous source, produced biosynthetically by one or more pathways, or derived from different sources in different compartments within the same cell. For example, in our model organism *Trypanosoma brucei*, inositol can be synthesized in the Golgi apparatus for glycosylphosphatidylinositol anchor synthesis or imported for use in bulk inositol lipid synthesis (Gonzalez-Salgado *et al.*, 2012).

The placement of metabolites in pathways is significantly improved by stable isotope tracing, which has recently elucidated some important descriptions of central metabolism (Lane *et al.*, 2011; Saunders *et al.*, 2011; Tang *et al.*, 2007; Zamboni and Sauer, 2009). For example, labeling patterns of aspartate and tricarboxylic acid (TCA) cycle intermediates in human lung cancer cells revealed anaplerosis via activation of pyruvate carboxylase (Fan *et al.*, 2009), while in yeast a novel riboneogenesis pathway was confirmed by metabolite labeling (Clasquin *et al.*, 2011). The major limitation with current isotope tracing approaches is the reliance on targeted analysis of labeled metabolites, making it impossible to trace precursor distribution into unexpected areas of metabolism. Nonetheless, these studies have already demonstrated that our classical view of metabolism in many organisms is overly simplistic, making an expansion of this approach beyond central metabolism highly desirable. An untargeted isotope tracing approach can be expected to reveal novel areas of metabolic networks that are essential to include in systems-based studies of metabolism.

Stable isotopes are becoming more commonly used in untargeted metabolomics studies to facilitate the quantification and identification of metabolites. Isotope labeled metabolite extracts can assist with an elemental formula assignment (Hegeman *et al.*, 2007; Wu *et al.*, 2005) or can provide internal standards for quantitative LC–MS-based metabolomics by isotope dilution (Wu *et al.*, 2005). Fully labeled extracts have also been applied to the differentiation of biogenic and exogenous features in LC–MS data (Bueschl *et al.*, 2012; Giavalisco *et al.*, 2009). Stable isotopes have been applied to targeted metabolomic studies for stable isotope tracing and, in combination with quantification, for flux analysis (Lane *et al.*, 2011; Tang *et al.*, 2009; Winder *et al.*, 2011). Expansion of these studies to the whole metabolic network requires accurate and efficient measurement of stable isotope distributions for all metabolites. Untargeted detection of isotopomer distributions has been demonstrated for gas chromatography–mass spectrometry (GC–MS); however, interpretation is difficult for unknown metabolites if they are not represented in spectral libraries (Hiller *et al.*, 2010; Kempa *et al.*, 2009). Here we introduce a method that

uses high-resolution LC–MS to perform untargeted detection of isotopic patterns for stable isotope labeled metabolites in complex mixtures and thus enables identification of novel metabolic pathways without preconception of the fate of labeled precursors.

10.1 Experimental section

10.1.1 Parasite culture, metabolite extraction and sample analysis

Labeling of cellular metabolites was achieved by growth of the procyclic-form *Trypanosoma brucei* under standard cell culture conditions in SDM-79 medium (Brun and Schönenberger, 1979), with 10 mM additional uniformly ^{13}C-labeled glucose (50% of total glucose labeled), for 5 days (10 doublings) to achieve close to steady-state labeling. A parallel culture was prepared under identical conditions, except with 10 mM additional unlabeled glucose to allow identification of unlabeled metabolites. A total of 108 parasites were quenched by rapid cooling to 0 °C (Saunders *et al.*, 2011), growth medium was removed after centrifugation, and metabolites extracted with a monophasic chloroform/methanol/water (1:3:1) mixture, which extracts both polar and nonpolar metabolite species, as previously described (Creek *et al.*, 2011). LC–MS analysis was performed with ZIC-HILIC hydrophilic interaction liquid chromatography (Merck Sequant), coupled to high-resolution Exactive Orbitrap mass spectrometry (Thermo, Hemel Hempstead, U.K.) operating in both positive and negative ionization modes, according to our published method (Creek *et al.*, 2011). Samples were analyzed in triplicate.

10.1.2 Metabolomics data processing

Initial data processing of raw LC–MS data (mzXML format) was performed using the standard IDEOM (`http://mzmatch.sourceforge.net/ideom.php`) workflow as described previously (Creek *et al.*, 2012a). This workflow utilizes the XCMS Centwave algorithm for peak detection (Tautenhahn *et al.*, 2008) and mzMatch.R for alignment of samples, filtering, and metadata storage in PeakML files (Scheltema *et al.*, 2011).Only metabolites detected in the unlabeled cell extracts were retained for further analysis. Parameters used for noise filtering and (putative) metabolite identification are available in Supplementary Table 1 in the Supporting Information.

10.1.3 Detection of stable isotope-labeled metabolites

The PeakML file after filtering and identification was scanned for labeled metabolites using the IDEOM Isotope Search and the PeakML.Isotope.UntargettedIsotope function of mzMatch-ISO (`http://mzmatch.sourceforge.net/isotopes-targetted.php`). The algorithm for detection of isotope labeling begins by first determining the number of possible labeled isotopomers a metabolite can have, according to its molecular formula or mass (e.g., $C_3H_6O_3$ could have 1, 2, or 3 carbons labeled). Then the expected mass of each labeled isotopomer is calculated based on the known mass difference between the light and heavy isotopes. Peaks are assigned as isotopomers if their mass is within 4 ppm (the mass accuracy of the spectrometer) of the predicted mass and within a retention time window of ±0.2 min (the average peak width) from the unlabeled peak, as isotopomers are expected to co-elute. This procedure is repeated for every possible isotopomer of each putatively annotated (unlabeled) metabolite to provide a comprehensive list of relative isotopomer intensities for each metabolite. If necessary, isotopomer signals that are missed in the initial processing, e.g., because of low intensities or irregular peak shapes, are gap-filled from raw data in a targeted manner by mzMatch-ISO. The accuracy and precision of the approach were validated by analysis of the natural ^{13}C isotope (M + 1.0033) in 60 identified metabolites in the unlabeled samples (n = 3). The measured relative isotope abundance was 1.1±1.3% lower than the theoretical isotope abundance (mean ± standard deviation for 60 metabolites), and the average standard deviation of experimental isotope abundances was 0.4% (range 0.005–1.8%).

Data are either visualized in IDEOM within Excel, showing the relative intensity of labeled isotopomers (see Supplementary File 1 (Excel spreadsheet) in the Supporting Information), or in mzMatch-ISO-generated PDF files containing detailed results for each metabolite. These results include raw chromatograms representing the monoisotopic and corresponding isotopic peaks; a normalized plot showing the variability in labeling between replicates; a trend plot of the pattern of labeling in each sample group; and a plot quantifying the absolute labeling pattern of a selected isotopomer of interest (see Supplementary File 2 (pdf file) in the Supporting Information). Both of these data outputs were collated to generate a comprehensive list of labeled metabolites. The procedure can be applied to ^{13}C, ^{15}N, ^{2}H, or ^{18}O isotope labels or user-defined mass differences, e.g., for chemical tags (Kasuya *et al.*, 2004).

10.2 Results and discussion

The method requires growth of the organism of interest in the presence of a predefined proportion (ideally 50%) of a stable isotope labeled nutrient. In our proof-of-concept example, we used procyclic-form *Trypanosoma brucei*, the protozoan parasite responsible for African sleeping sickness, which due to its reduced and well studied metabolism is a suitable model organism for validating our approach (Creek *et al.*, 2012b). Growth medium containing 50% uniformly ^{13}C-labeled glucose as the main carbon source was used to generate steady-state labeling of the cellular metabolome. A parallel culture was incubated with unlabeled glucose at the same concentration. Quenching and metabolite extraction was achieved by a nonspecific metabolite extraction protocol, and LC–MS analysis utilized zwitterionic hydrophilic interaction liquid chromatography (ZIC-HILIC) coupled to high-resolution mass spectrometry using Orbitrap technology (Creek *et al.*, 2011). Data were analyzed using a combination of freely available software, XCMS (Smith *et al.*, 2006), mzmatch.R (Scheltema *et al.*, 2011), and IDEOM (Creek *et al.*, 2012a), with customized additions to detect and quantify isotopomer signals.

In our example, over 27,000 unique peak groups (defined by m/z ratio and retention time) were detected in both positive and negative mode ionization. The vast majority of these peaks are noise or artifacts arising from the mass spectrometry, chromatography, sample preparation, and data processing (see Supplementary Figure 1 in the Supporting Information) (Brown *et al.*, 2011; Creek *et al.*, 2011). A key component of this method is application of the default noise filters of mzMatch.R and IDEOM to obtain a list of monoisotopic peaks representing putative metabolites present in the label-free sample. A total of 82 metabolites were identified confidently by exact mass and retention time based on authentic metabolite standards (level 1 identification according to the Metabolomics Standards Initiative (MSI) (Sumner *et al.*, 2007)). A total of 506 putative annotations (levels 2 and 3, MSI) were made using exact mass and predicted retention times from the IDEOM database (see Supplementary File 1 (Excel spreadsheet) in the Supporting Information) (Creek *et al.*, 2011, 2012a).

The extent and pattern of labeling in each (putative) metabolite was determined by an automated search of all potential isotopomers according to accurate mass and retention time. Comparison of isotope distributions with unlabeled samples, and with theoretical natural isotope abundances, confirmed the presence of stable isotope labeling for 187 metabolites (Figure 10.1). Visualization of labeled metabolites in the global metabolic network (Yamada *et al.*, 2011) reveals incorporation of glucose-

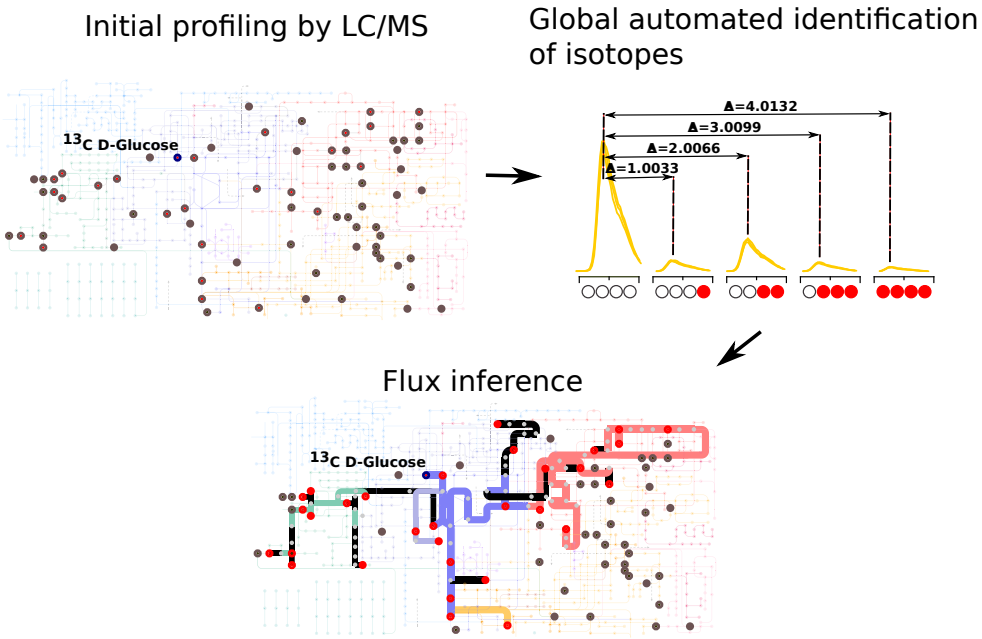


Figure 10.1: Schematic representation of the method. Red dots represent labeled metabolites.

derived carbon into diverse areas of metabolism (Figure 10.2).

The isotope labeling patterns generated from 50% U-^{13}C-labeled glucose allowed immediate identification of many active metabolic pathways within the cell. For example, the distinct labeling pattern for succinate (Figure 10.3a) of predominantly three labeled carbons is consistent with previous targeted studies of trypanosomes that demonstrated a fermentation pathway primarily responsible for succinate production, rather than a traditional TCA cycle source (Coustou *et al.*, 2008; van Weelden *et al.*, 2003). The results also show a similar labeling pattern in aspartate and orotate, confirming biosynthesis of aspartate from phosphoenolpyruvate (via oxaloacetate) and subsequent *de novo* pyrimidine synthesis. The enzymes required for *de novo* pyrimidine synthesis have been identified and studied in bloodstream form trypanosomes (Arakaki *et al.*, 2008; Hammond and Gutteridge, 1982) and here we provide the first direct evidence for *de novo* pyrimidine synthesis from glucose-derived aspartate in the procyclic life-cycle stage despite the high abundance of aspartate and pyrimidines available in the growth medium.

Extensive labeling was observed in nucleotides and lipids (Figure 10.2 and Sup-

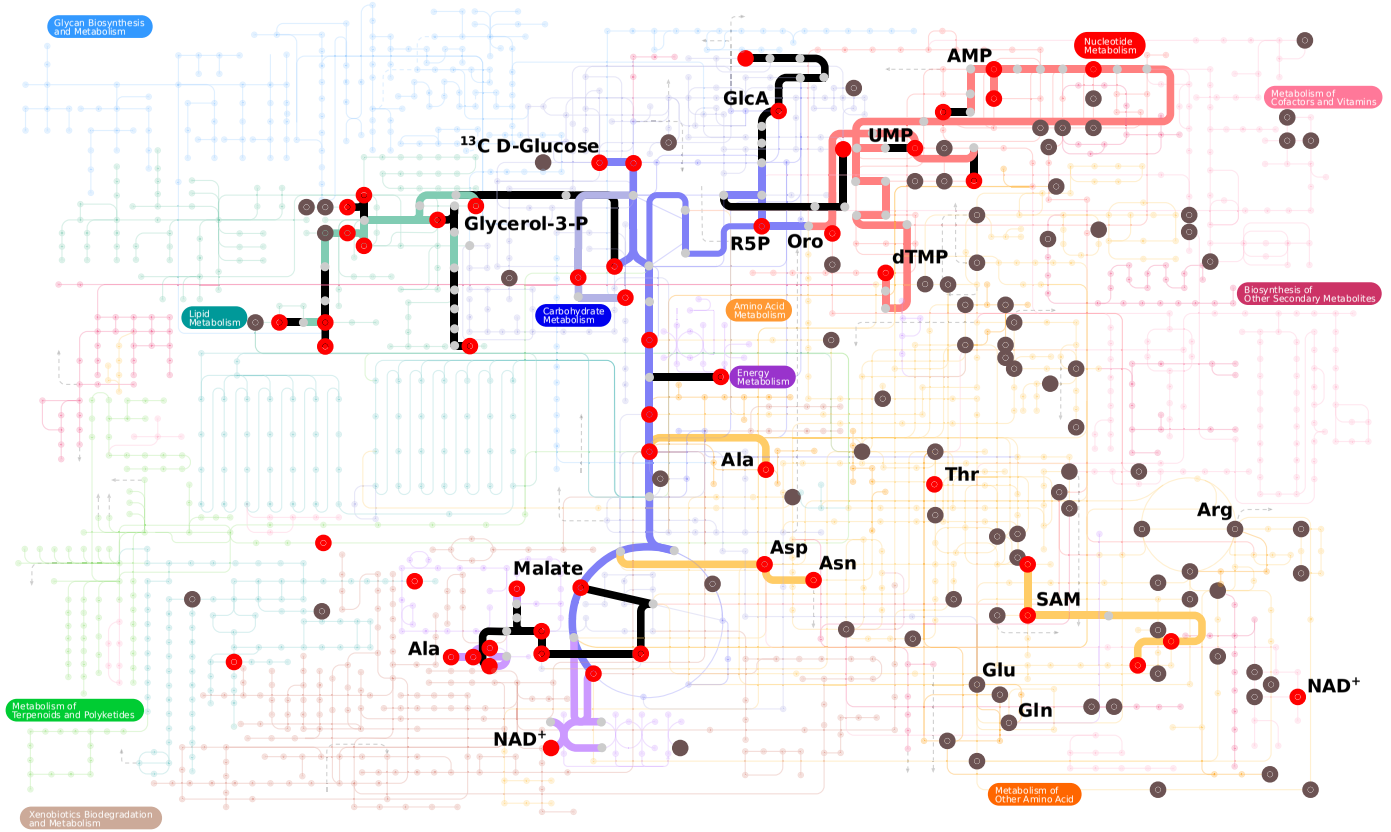


Figure 10.2: KEGG global pathway map of identified and putatively annotated metabolites labeled by ^{13}C-glucose (red) and unlabeled putative metabolites arising from alternative sources (gray). Inferred potentially active biosynthesis pathways are highlighted; black routes have not previously been known to be active in Trypanosoma brucei. Map generated using IPath tools (Yamada *et al.*, 2011).

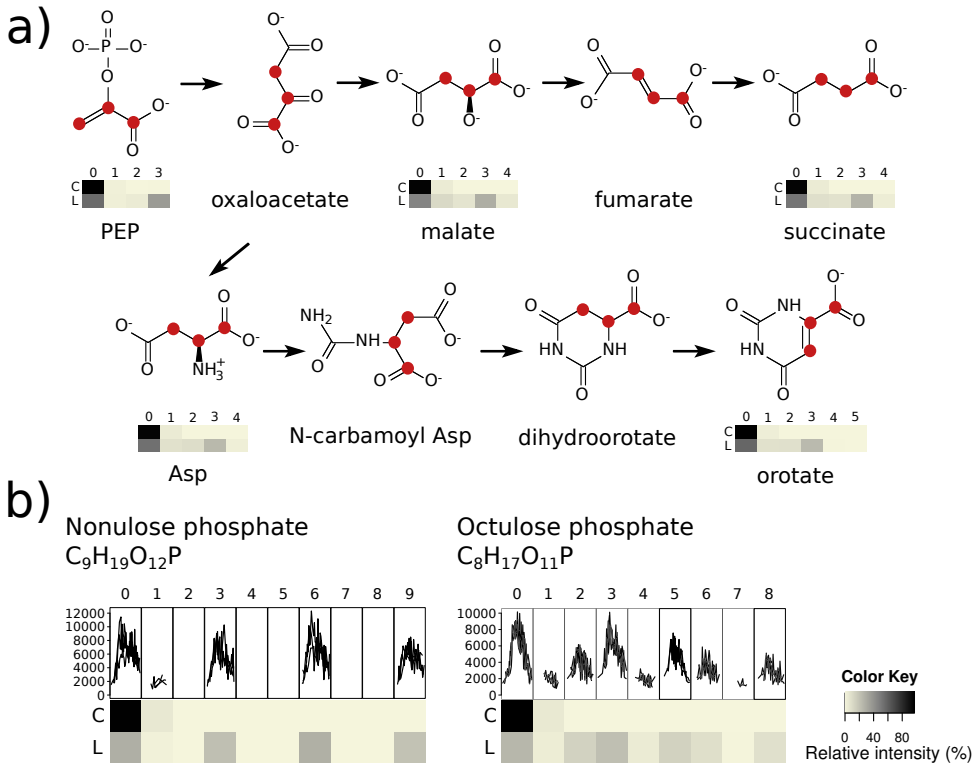


Figure 10.3: (**a**) Labeling pattern in succinate, aspartate, and orotate biosynthesis. In contrast to the expected predominant 2- and 4-carbon labeling from the TCA cycle, the majority of molecules contain three labeled carbons, consistent with a fermentative source of succinate. (**b**) Heavy-isotope labeling patterns in the new trypanosomal metabolites putatively annotated as octulose phosphate and nonulose phosphate. C represents the unlabeled control, and L represents the ^{13}C-glucose labeled sample.

plementary File 1 (Excel spreadsheet) in the Supporting Information). Nucleotide labeling patterns confirm both *de novo* pyrimidine synthesis (*vide supra*) and active purine salvage pathways incorporating ribose 5-phosphate (five labeled carbons) from the pentose phosphate pathway (PPP), while glycerophospholipid labeling patterns revealed incorporation of three carbons from glycerol 3-phosphate or dihydroxyacetone phosphate (products of glycolysis) and also two carbon lipid chain extensions from acetyl-CoA (Smith and Bütikofer, 2010). While acetyl-CoA was not observed in this study (being incompatible with this chromatographic method), its presence can be inferred from the two-carbon incorporation into lipids and by the presence of

two-carbon labeled acetyllysine and acetylcarnitine, although the relatively low level of label incorporation into these metabolites (15% and 21%, respectively) confirms previous reports that glucose is not the primary source of acetyl-CoA (Rivière *et al.*, 2009). A single LC column that optimally separates polar and nonpolar metabolites has yet to be successfully employed in metabolomics research. Notwithstanding, our nontargeted approach is clearly capable of identifying metabolites from a broad physicochemical spectrum and can be readily adapted to protocols involving multiple columns to separate metabolites if even more comprehensive coverage is required.

A key feature of the untargeted metabolomics approach is the discovery of unexpected labeling in metabolites and of labeling in unexpected metabolites. The most striking cases of unexpected labeled metabolites detected in our case study included glycerate and gluconate, potentially arising from dephosphorylation of glycolysis-derived phosphoglycerate and PPP-derived phosphogluconate, respectively, although specific enzymes for these reactions have not been identified in *T. brucei*. Importantly, prior knowledge of putative metabolite identities is not essential for detection of isotopomers with our method, and isotopomer profiling of unidentified peaks facilitates the discovery of novel metabolites. In this example, two labeled metabolites that were not present in online metabolite databases were detected with m/z, retention time, and labeling patterns consistent with octulose phosphate and nonulose phosphate (Figure 10.3b). These metabolites are highly relevant for trypanosome biology as they indicate the presence of new links between the glycolytic and pentose phosphate pathways analogous to recent findings in yeast (Clasquin *et al.*, 2011), and further work on these novel pathways may enable the discovery of new targets for trypanosomal drugs.

10.3 Conclusions

As we show, untargeted LC–MS-based metabolomics is able to successfully map the extent of metabolism of stable isotope-labeled glucose in procyclic-form trypanosomes. The results, obtained in a single experiment, are generally consistent with many years of prior research and genome annotations and also highlight new areas of metabolism that would not have been detected by a traditional targeted approach. The necessary computational analysis is straightforward and has been automated and included in the mzMatch.R and IDEOM applications (`http://mzmatch.sourceforge.net`), making it easy to apply the same strategy to any organism to confirm the presence of metabolic pathways, discover new pathways, confirm metabolic activity under specific growth conditions or, by inclusion of multiple samples

at specified time points with concurrent quantitative analysis, measure flux through a network.

Acknowledgements

D.J.C. is supported by an NHMRC postdoctoral training fellowship. Metabolomics analysis was provided by the Scottish Metabolomics Facility. This work was partly supported by the Wellcome Trust through The Wellcome Trust Centre for Molecular Parasitology, which is supported by core funding from the Wellcome Trust (Grant 085349).

Supporting information

Supplementary Figure 1, distribution of sources of noise peaks removed by automated filtering; Supplementary Table 1, table of parameters applied for data processing with XCMS, mzMatch, and IDEOM; Supplementary File 1, Excel spreadsheet summary of identified and putatively annotated metabolites in procyclic-form *Trypanosoma brucei*, ^{13}C-labeled metabolites are annotated in column J, detailed peak intensities for each metabolite and isotopomer are listed in the "RAWisotopomer-PEAKS" sheet (All data from this study, and full functionality for data analysis and visualization in Excel, are available in the IDEOM file (`http://puma.ibls.gla.ac.uk/untargeted/IDEOM_13Cglucose_PCF.xlsb`); and Supplementary File 2, PDF file containing all mass traces and graphs for isotopomers of identified and annotated metabolites. This material is available free of charge via the Internet at `http://pubs.acs.org/doi/suppl/10.1021/ac3018795`.

Chapter 11

Conclusions and future perspectives

Parts of this chapter were published in:

M. Elena Merlo[1,2], Andris Jankevics[1,3], Eriko Takano[2], Rainer Breitling[1,3]

Bioanalysis 3(21):2443-58, 2011.

Maya Berg[4], Manu Vanaerschot[4], Andris Jankevics[1,3,5], Bart Cuypers[1], Rainer Breitling[1,3,5], Jean-Claude Dujardin[4,6]

Comput. Struct. Biotech. J. 4(5):e201301002, 2013

Quoc-Thai Nguyen [1,2], M. Elena Merlo[1,2], Marnix H. Medema[1,2], Andris Jankevics[1,3], Rainer Breitling[1,3], Eriko Takano[2]

FEBS Letters 586(15):2177-83, 2012.

1 Groningen Bioinformatics Center, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Nijenborgh 7, 9747 AG Groningen, The Netherlands

2 Microbial Physiology, Groningen Biomolecular Sciences & Biotechnology Institute, University of Groningen, Nijenborgh 7, 9747 AG Groningen, The Netherlands

3 Institute of Molecular, Cell & Systems Biology, College of Medical, Veterinary & Life Sciences, University of Glasgow, Joseph Black Building B3.10, Glasgow, G12 8QQ, UK

4 Unit of Molecular Parasitology, Department of Biomedical Sciences, Institute of Tropical Medicine, Nationalestraat 155, 2000 Antwerp, Belgium

5 Faculty of Life Sciences, Manchester Institute of Biotechnology, University of Manchester, 131 Princess Street, Manchester M1 7DN, UK

6 Department of Biomedical Sciences, University of Antwerp, Universiteitsplein 1, 2610 Antwerp, Belgium

abstract

The improvement in technology in the near future will undoubtedly lead to a new generation of mass spectrometers with even higher resolving power and mass accuracy – and with a further explosion of the amount of data generated. We expect that the development of analytical platforms will have to go hand-in-hand with the evolution of the available software tools, and that as a result we will achieve a significantly increased coverage of the metabolome in untargeted experiments. Although the developed approaches presented in this thesis led to a greater understanding of the metabolome of interest, there is a still lot of space for improvement.

Metabolomics has made giant steps in the last decade, thanks to the availability of an increasing number of effective procedures for sample preparation and highly accurate metabolite detection techniques as well as the vast progress in data processing and visualization. Despite these advances, the detected metabolite cover only a part of the predicted in silico metabolomes for microbes and often the secondary metabolism is not represent it at all. So the question is: what can we do next? Can we still aim to cover all metabolites produced by a cell?

Several methods coupling chromatography to MS are available and were usually developed and optimized for targeted analysis of specific classes of compounds. For untargeted metabolomics, metabolite detection could be expanded by the use of orthogonal techniques starting from the application of diversified quenching and extraction techniques during sample preparation, to complementary separation techniques (e.g., hydrophilic and reverse phase liquid chromatography), as well as a combination of analytical platforms (e.g., GC and LC). Already in the sample preparation phase there is still much room for improvement by specifically targeting those classes of compounds that are usually not recovered with the general extraction protocols.

LC–MS platform could further benefit from technical improvements for more stable column design (to minimize inter-column variation) while awaiting the next-generation mass spectrometers with even higher resolving power, mass accuracy and multiple fragmentation techniques. Although the data processing step is becoming more accessible to non-specialists, future software development efforts will need to focus on making the computational pipeline more user friendly, without sacrificing performance, and to obtain a better match between the computational output and the interpretation that is achievable by manual assessment by an experienced analytical chemist. Last but not least, the increase of large-scale metabolomics projects urges the further development of normalization methods to pool results of different samples that need to be compared.

It is clear that studying the metabolome, lying closest to the phenotype, together with other global molecular profiles, such as the genome, transcriptome or proteome, can significantly enhance our insights into the interactions between the different components of biological systems and how these interactions give rise to a specific behavior of that system and result in a phenotype (Snoep and Westerhoff, 2005). Furthermore, the integration of metabolomics and genomics datasets will also contribute to differentiate 'driver' mutations from biologically neutral 'passenger' changes. However, the integration of different molecular profiling datasets into one comprehensive, easily consultable entity requires an even greater deal of bioinformatics. Even in arcane areas of biology, such as the research on protozoan parasites causing neglected tropical diseases, a rich tradition of metabolomics research has accumulated in a surprisingly short time (Breitling *et al.*, 2012; Canuto *et al.*, 2012; Creek *et al.*, 2012b; De Souza *et al.*, 2006; Imbert *et al.*, 2012; Kamleh *et al.*, 2008; Saunders *et al.*, 2010, 2011; Scheltema *et al.*, 2010; Silva *et al.*, 2011; t'Kindt *et al.*, 2010a,b; Vincent *et al.*, 2012; Zheng *et al.*, 2010). Untargeted metabolomics studies have shown to disclose complete pathways that respond to drug action, contributing to unraveling the mode of action of these drugs (Vincent *et al.*, 2012). We expect that even deeper insights into more complex phenotypes, especially the vexing issue of emerging drug resistance, will be provided by carefully designed metabolomics studies in the coming years. Integration of these data into "Silicon Parasites", i.e. comprehensive, multi-scale models of protozoan physiology will play a key role in selecting ideal drug targets for developing new anti-parasite drugs.

It is essential to keep in mind that metabolomics itself cannot provide a comprehensive understanding of how metabolic systems work, as it only provides steadystate snapshots of metabolism at the studied experimental condition. Besides, providing a biological interpretation of metabolomics results is not conceivable without locating them in the context of a biological network. To explore the dynamics of metabolite levels and the causal mechanisms leading to metabolite depletion and accumulation, additional analyses are necessary.

In a systems biology context, metabolite profiles can be integrated with the information collected from sequenced genomes, transcriptomic and proteomic data, as well as with metabolic flux measurements and genome-scale model reconstructions. Only by the combination of experimental approaches with quantitative *in silico* models we can truly attempt to understand the behavior of microorganisms in their diverse living environments. Two recent reviews have analyzed the benefits of integrating different disciplines in order to generate a systems-level understanding of microbial metabolism. Reaves and Rabinowitz (2011) highlighted recent integrated

studies which have successfully identified novel metabolic pathways and provided a better understanding of regulatory and homeostasis mechanisms. Heinemann and Sauer (2010) analyzed the different systems biology approaches with the conclusion that computational methods and mathematic modeling have a primary role in this multidisciplinary area, to achieve a better interpretation of the existing biological data but also to efficiently guide future experiments.

Genome-wide metabolic models for different microorganisms have been reconstructed from genome annotations and refined with experimental data available from the literature and comparative genomics (Feist *et al.*, 2009; Price *et al.*, 2004). These *in silico* stoichiometric networks describe the stoichiometric constraints on flux distribution in a metabolic system and can be used to predict the response of cellular metabolism to perturbations (especially gene knock-out) using computational tools such as flux balance analysis (Orth *et al.*, 2010). Genomic information can be used to generate additional regulatory constraints by indicating group of genes that tend to be on/off under similar genetic or environmental conditions (Park *et al.*, 2010).

The predicted fluxes are usually validated on well-studied pathways as for example pathways of core metabolism for which comprehensive genomic, transcriptomic and stoichiometric data are available. In the context of metabolomics, it will be particularly interesting to expand the scope of computational modeling to explore the organism-specific fringes of the metabolome (e.g., the production of secondary metabolites) (Breitling *et al.*, 2008). This will expand the usefulness of microbial metabolomics as a diagnostic tool in synthetic microbiology (Medema *et al.*, 2011). Furthermore, the unexplored metabolome of many microbes is likely to contain compounds with interesting bioactivities and therapeutic potential, for example as antibiotics (Davies, 2010), which could be discovered using metabolomics. Computational modeling based on flux balance analysis allows evaluating a priori which modification of biochemical pathways could enhance metabolic fluxes towards a compound of interest (e.g., a secondary metabolite with high commercial value) — and after the corresponding designer microbe has been generated, metabolomics can be used to identify pleiotropic effects as well as unexpected metabolic bottlenecks, such as the depletion of precursors or accumulation of toxic side products.

The application of metabolomics in the "debugging" of engineered microbial strains is not limited to the analysis of the pathway of interest. It is often even more important to characterize unexpected pleiotropic effects on the system. In **Chapter 9** we have used untargeted metabolomics to characterize the global metabolic rearrangement following induction of a non-coding antisense RNA targeting glutamine synthetase I in *Streptomyces coelicolor*: we could show that this very specific manip-

ulation resulted in a "synthetic metabolic switch" with widespread and rapid changes of metabolite levels. Considering that the natural biosynthesis of secondary metabolites is often the result of stress conditions and accompanied by a metabolic switch (see, e.g. (Nieselt *et al.*, 2010)), this is important background information for future attempts at awakening biosynthetic gene clusters. The same holds true for the systematic, unbiased characterization of the metabolomic response to general stress conditions (e.g., salt stress (Kol *et al.*, 2010)), which are known to induce or favor the overproduction of secondary metabolites in a natural environment.

In general, MS-based metabolomics represents the best available tool for validation of the predictions given by computational metabolic modeling and in the discovery of novel biochemical routes. For example, Mo *et al.* (2009) demonstrated how extracellular metabolome measurements can be integrated with genome-scale metabolic network models to correctly predict variation in intracellular metabolite levels in *S. cerevisiae* during ammonium assimilation. In another study by Woo *et al.* (2010), untargeted profiling of intracellular metabolites in *Corynebacterium glutamicum* led to the discovery that this well-studied organism responds to phosphate starvation with an unexpected accumulation of glycogen; this finding was in agreement with the predicted in silico response of the genome-scale model of *C. glutamicum* to which missing reactions for glycogen synthesis and degradation had been added manually. In yeast, comparative metabolomics guided towards the characterization of the protein SHB17, for which gene sequence and expression data were available but whose function was still unknown. The differences in metabolite profiles between an SHB17 knockout and the wild type strain highlighted the accumulation of a few compounds putatively used as substrates by the enzyme. Further biochemical assays and elucidation of the protein crystal structure revealed that SBH17 is involved in riboneogenesis, an alternative route to produce ribose-5-phosphate from intermediates of glycolysis without production of NADPH (Clasquin *et al.*, 2011). In addition to flux balance analysis, processing of metabolomics data based on correlation analysis of the detected metabolites in combination with genetical genomics approaches have been proved to be successful for pathway reconstructions (Breitling *et al.*, 2008; Keurentjes *et al.*, 2006).

Metabolomics has an important role in the elucidation and in the confirmation of novel biochemical pathways, especially when used in combination with ^{13}C-labeling experiments. For example, Peyraud *et al.* (2009) used this strategy to demonstrate that the methylotrophic bacterium *Methylobacterium extorquens* AM1 uses the ethylmalonyl-CoA pathway for the assimilation of one-carbon substrates and regeneration of glyoxylate. The authors utilized LC–Orbitrap MS to detect the pres-

ence of specific CoA substrates and to discriminate between two different proposed biochemical pathways. ^{13}C-labeling metabolomics was used to further confirm the proposed biochemical steps of the ethylmalonyl-CoA pathway used for the synthesis of glyoxylate. In addition, NMR measurements of the positional isotopomers of glycine helped to evaluate the contribution of this newly discovered pathway to glyoxylate regeneration in comparison with other central metabolic pathways, which are also able to synthesize glyoxylate. As a continuation on this topic, Okubo *et al.* (2010) discovered an alternative biochemical route in the utilization of glyoxylate by combination of microarray analysis, phenotype screening, MS-based metabolite measurements and ^{13}C-labeling experiments. The study led to the discovery that the bacterium can use a second pathway to convert glyoxylate to 2-phosphoglycerate via glycine and part of the serine cycle. Furthermore, it was demonstrated that the function of the enzyme mclA1, which has an important role in the ethylmalonyl-CoA pathway, can be complemented by a second malyl-CoA/betamethylmalyl-CoA lyase, MclA2. These studies emphasize the role that MS- and NMR-based platforms, when used in combination with isotope-labeling techniques, may have in the characterization of novel biochemical pathways and alternative enzymatic functions (reviewed by Dauner (2010) and by Zamboni and Sauer (2009)).

The complementary use of metabolite analytics, fluxomics and computational modelling is essential for progress in microbial metabolomics, in particular as a tool for the ambitious projects of synthetic biology, which aims to design microbial systems *de novo*. As a key element in the bioanalytical toolbox, metabolomics will play an increasing role in the refinement and functional diagnosis of newly engineered microbial strains.

Bibliography

Alam M. T., Merlo E. M., The STREAM Consortium, Hodgson D. A., Wellington E. M. H., Takano E., and Breitling R. Metabolic modeling and analysis of the metabolic switch in *Streptomyces coelicolor*. *BMC Genomics* **11**, 202, 2010.

Ali N., Herron P. R., Evans M. C., and Dyson P. J. Osmotic regulation of the *Streptomyces lividans* thiostrepton-inducible promoter, *ptipA*. *Microbiology* **148**(2), 381–390, 2002.

Amon J., Titgemeyer F., and Burkovski A. Common patterns - unique features: nitrogen metabolism and regulation in Gram-positive bacteria. *FEMS Microbiol. Rev.* **34**(4), 588–605, 2010.

Annesley T. M. Ion suppression in mass spectrometry. *Clin. Chem.* **49**(7), 1041–1044, 2003.

Arakaki T. L., Buckner F. S., Gillespie J. R., Malmquist N. A., Phillips M. A., Kalyuzhniy O., Luft J. R., Detitta G. T., Verlinde C. L., Van Voorhis W. C., Hol W. G., and Merritt E. A. Characterization of *Trypanosoma brucei* dihydroorotate dehydrogenase as a possible drug target; structural, kinetic and RNAi studies. *Mol. Microbiol.* **68**(1), 37–50, 2008.

Arbona V., Iglesias D. J., Talón M., and Gómez-Cadenas A. Plant phenotype demarcation using nontargeted LC–MS and GC–MS metabolite profiling. *J. Agric. Food Chem.* **57**(16), 7338–7347, 2009.

Arita M. What can metabolomics learn from genomics and proteomics? *Curr. Opin. Biotechnol.* **20**(6), 610–615, 2009.

Benjamini Y. and Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. B* **57**(1), 289–300, 1995.

Bentley S. D., Chater K. F., Cerdeño-Tárraga A. M., Challis G. L., Thomson N. R., James K. D., Harris D. E., Quail M. A., Kieser H., Harper D., Bateman A., Brown S., Chandra G., Chen C. W., Collins M., Cronin A., Fraser A., Goble A., Hidalgo J., Hornsby T., Howarth S., Huang C.-H., Kieser T., Larke L., Murphy L., Oliver K., O'Neil S., Rabbinowitsch E., Rajandream M.-A., Rutherford K., Rutter S., Seeger K., Saunders D., Sharp S., Squares R., Squares S., Taylor K., Warren T., Wietzorrek A., Woodward J., Barrell B. G., Parkhill J., and Hopwood D. A. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* **417**(6885), 141–147, 2002.

Betts J. I. and Baganz F. Miniature bioreactors: current practices and future opportunities. *Microb. Cell Fact.* **5**, 21, 2006.

Blekherman G., Laubenbacher R., Cortes D. F., Mendes P., Torti F. M., Akman S., Torti S. V., and Shulaev V. Bioinformatics tools for cancer metabolomics. *Metabolomics* **7**(3), 329–343, 2011.

Boccard J., Veuthey J. L., and Rudaz S. Knowledge discovery in metabolomics: an overview of MS data handling. *J. Sep. Sci.* **33**(3), 290–304, 2010.

Bode H. B., Bethe B., Höfs R., and Zeeck A. Big effects from small changes: possible ways to explore nature's chemical diversity. *Chembiochem* **3**(7), 619–627, 2002.

Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* **32**(90001), D267–270, 2004.

Bölling C. and Fiehn O. Metabolite profiling of *Chlamydomonas reinhardtii* under nutrient deprivation. *Plant Physiol.* **139**(4), 1995–2005, 2005.

Bolten C. J., Kiefer P., Letisse F., Portais J.-C., and Wittmann C. Sampling for metabolome analysis of microorganisms. *Anal. Chem.* **79**(10), 3843–3849, 2007.

Böttcher C., Roepenack-Lahaye E. v., Willscher E., Scheel D., and Clemens S. Evaluation of matrix effects in metabolite profiling based on capillary liquid chromatography electrospray ionization quadrupole time-of-flight mass spectrometry. *Anal. Chem.* **79**(4), 1507–1513, 2007 PMID: 17297948.

Breitling R., Ritchie S., Goodenowe D., Stewart M. L., and Barrett M. P. *Ab initio* prediction of metabolic networks using fourier transform mass spectrometry data. *Metabolomics* **2**(3), 155–164, 2006a.

Breitling R., Pitt A. R., and Barrett M. P. Precision mapping of the metabolome. *Trends Biotechnol.* **24**(12), 543–548, 2006b.

Breitling R., Vitkup D., and Barrett M. P. New surveyor tools for charting microbial metabolic maps. *Nat. Rev. Microbiol.* **6**(2), 156–161, 2008.

Breitling R., Bakker B. M., Barrett M. P., Decuypere S., and Dujardin J.-C. Metabolomic systems biology of protozoan parasites. In K. Suhre, editor, *Genetics Meets Metabolomics* pages 73–84. Springer New York, 2012.

Brown M., Dunn W. B., Dobson P., Patel Y., Winder C. L., Francis-McIntyre S., Begley P., Carroll K., Broadhurst D., Tseng A., Swainston N., Spasic I., Goodacre R., and Kell D. B. Mass spectrometry tools and metabolite-specific databases for molecular identification in metabolomics. *Analyst* **134**(7), 1322–1332, 2009.

Brown M., Wedge D. C., Goodacre R., Kell D. B., Baker P. N., Kenny L. C., Mamas M. A., Neyses L., and Dunn W. B. Automated workflows for accurate mass-based putative metabolite identification in LC/MS-derived metabolomic datasets. *Bioinformatics* **27**(8), 1108–1112, 2011.

Brun R. and Schönenberger M. Cultivation and *in vitro* cloning or procyclic culture forms of *Trypanosoma brucei* in a semi-defined medium. Short communication. *Bioinformatics* **36**(3), 289–292, 1979.

Büchs J. Introduction to advantages and problems of shaken cultures. *Biochem. Eng. J.* **7**(2), 91–98, 2001.

Bueschl C., Kluger B., Berthiller F., Lirk G., Winkler S., Krska R., and Schuhmacher R. MetExtract: A new software tool for the automated comprehensive extraction of metabolite-derived LC/MS signals in metabolomics research. *Bioinformatics* **28**(5), 736–738, 2012.

Canuto G. A., Castilho-Martins E. A., Tavares M., López-Gonzálvez Á., Rivas L., and Barbas C. CE–ESI–MS metabolic fingerprinting of *Leishmania* resistance to antimony treatment. *Electrophoresis* **33**(12), 1901–1910, 2012.

Caspi R., Foerster H., Fulcher C. A., Kaipa P., Krummenacker M., Latendresse M., Paley S., Rhee S. Y., Shearer A. G., Tissier C., Walk T. C., Zhang P., and Karp

P. D. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **36**(Database issue), D623–631, 2008.

Caspi R., Altman T., Dale J. M., Dreher K., Fulcher C. A., Gilham F., Kaipa P., Karthikeyan A. S., Kothari A., Krummenacker M., Latendresse M., Mueller L. A., Paley S., Popescu L., Pujar A., Shearer A. G., Zhang P., and Karp P. D. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **38**(Database issue), D473–479, 2010.

Challis G. L. and Hopwood D. A. Synergy and contingency as driving forces for the evolution of multiple secondary metabolite production by *Streptomyces* species. *Proc. Natl. Acad. Sci.* **100**(Suppl 2), 14555–14561, 2003.

Chaneton B., Hillmann P., Zheng L., Martin A. C. L., Maddocks O. D. K., Chokkathukalam A., Coyle J. E., Jankevics A., Holding F. P., Vousden K. H., Frezza C., O'Reilly M., and Gottlieb E. Serine is a natural ligand and allosteric activator of pyruvate kinase M2. *Nature* **491**(7424), 458–462, 2012.

Chappuis F., Sundar S., Hailu A., Ghalib H., Rijal S., Peeling R. W., Alvar J., and Boelaert M. Visceral leishmaniasis: what are the needs for diagnosis, treatment and control? *Nat. Rev. Microbiol.* **5**(11), 873–882, 2007.

Chavali A. K., Whittemore J. D., Eddy J. A., Williams K. T., and Papin J. A. Systems analysis of metabolism in the pathogenic trypanosomatid *Leishmania major*. *Mol. Syst. Biol.* **4**, 177, 2008.

Chen L. and Vitkup D. Distribution of orphan metabolic activities. *Trends Biotechnol.* **25**(8), 343–348, 2007.

Christin C., Smilde A. K., Hoefsloot H. C. J., Suits F., Bischoff R., and Horvatovich P. L. Optimized time alignment algorithm for LC–MS data: Correlation optimized warping using component detection algorithm-selected mass chromatograms. *Anal. Chem.* **80**(18), 7012–7021, 2008.

Chukualim B., Peters N., Fowler C., and Berriman M. Trypanocyc — a metabolic pathway database for *Trypanosoma brucei*. *BMC Bioinformatics* **9**(Suppl 10), P5, 2008.

Clardy J., Fischbach M. A., and Walsh C. T. New antibiotics from bacterial natural products. *Nat. Biotechnol.* **24**(12), 1541–1550, 2006.

Clasquin M. F., Melamud E., Singer A., Gooding J. R., Xu X., Dong A., Cui H., Campagna S. R., Savchenko A., Yakunin A. F., Rabinowitz J. D., and Caudy A. A. Riboneogenesis in yeast. *Cell* **145**(6), 969–980, 2011.

Coustou V., Biran M., Breton M., Guegan F., Riviére L., Plazolles N., Nolan D., Barrett M. P., Franconi J. M., and Bringaud F. Glucose-induced remodeling of intermediary and energy metabolism in procyclic *Trypanosoma brucei*. *J. Biol. Chem.* **283**(24), 16342–16354, 2008.

Creek D. J., Jankevics A., Breitling R., Watson D. G., Barrett M. P., and Burgess K. E. V. Towards global metabolomics analysis with liquid chromatography/mass spectrometry: improved metabolite identification by retention time prediction. *Anal. Chem.* **83**(22), 8703–8710, 2011.

Creek D. J., Jankevics A., Burgess K. E. V., Breitling R., and Barrett M. P. IDEOM: an Excel interface for analysis of LC–MS-based metabolomics data. *Bioinformatics* **28**(7), 1048–1049, 2012a.

Creek D. J., Anderson J., McConville M. J., and Barrett M. P. Metabolomic analysis of trypanosomatid protozoa. *Mol. Biochem. Parasitol.* **181**(2), 73–84, 2012b.

Creek D. J., Chokkathukalam A., Jankevics A., Burgess K. E., Breitling R., and Barrett M. P. Stable isotope-assisted metabolomics for network-wide metabolic pathway elucidation. *Anal. Chem.* **84**(20), 8442–8447, 2012c.

Cubbon S., Antonio C., Wilson J., and Thomas-Oates J. Metabolomic applications of HILIC–LC–MS. *Mass Spectrom. Rev.* **29**(5), 671–684, 2010.

Dai Y., Li Z., Xue L., Dou C., Zhou Y., Zhang L., and Qin X. Metabolomics study on the anti-depression effect of xiaoyaosan on rat model of chronic unpredictable mild stress. *J. Ethnopharmacol.* **128**(2), 482–489, 2010.

D'Alia D., Nieselt K., Steigele S., Müller J., Verburg I., and Takano E. Noncoding RNA of glutamine synthetase I modulates antibiotic production in *Streptomyces coelicolor* A3(2). *J. Bacteriol.* **192**(4), 1160–1164, 2010.

Dauner M. From fluxes and isotope labeling patterns towards in silico cells. *Curr. Opin. Biotechnol.* **21**(1), 55–62, 2010.

Davies J. The garden of antimicrobial delights. *F1000 Biol Rep* **2**, 2010.

De Souza D. P., Saunders E. C., McConville M. J., and Likić V. A. Progressive peak clustering in GC–MS metabolomic experiments applied to *Leishmania* parasites. *Bioinformatics* **22**(11), 1391–1396, 2006.

De Vos R. C., Moco S., Lommen A., Keurentjes J. J., Bino R. J., and Hall R. D. Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry. *Nat. Protoc.* **2**(4), 778–791, 2007.

Decuypere S., Vanaerschot M., Rijal S., Yardley V., Maes L., de Doncker S., Chappuis F., and Dujardin J.-C. Gene expression profiling of *Leishmania (Leishmania) donovani*: overcoming technical variation and exploiting biological variation. *Parasitology* **135**(2), 183–194, 2008.

Dettmer K., Aronov P. A., and Hammock B. D. Mass spectrometry-based metabolomics. *Mass Spectrom. Rev.* **26**(1), 51–78, 2007.

Deutsch E. W. Mass spectrometer output file format mzML. In *Proteome Bioinformatics* volume 604 of *Methods in Molecular Biology* pages 319–331. Humana Press, 2010.

Downing T., Imamura H., Decuypere S., Clark T. G., Coombs G. H., Cotton J. A., Hilley J. D., de Doncker S., Maes I., Mottram J. C., Quail M. A., Rijal S., Sanders M., Schonian G., Stark O., Sundar S., Vanaerschot M., Hertz-Fowler C., Dujardin J.-C., and Berriman M. Whole genome sequencing of multiple *Leishmania donovani* clinical isolates provides insights into population structure and mechanisms of drug resistance. *Genome Res.* **21**(12), 2143–2156, 2011.

Doyle M., MacRae J., De Souza D. P., Saunders E., McConville M., and Likić V. LeishCyc: a biochemical pathways database for *Leishmania major*. *BMC Syst. Biol.* **3**(1), 57, 2009.

Dunn W. B. Current trends and future requirements for the mass spectrometric investigation of microbial, mammalian and plant metabolomes. *Phys. Biol.* **5**(1), 011001, 2008.

Dunn W. B., Broadhurst D., Ellis D. I., Brown M., Halsall A., O'Hagan S., Spasic I., Tseng A., and Kell D. B. A GC–TOF–MS study of the stability of serum and urine metabolomes during the UK Biobank sample collection and preparation protocols. *Int. J. Epidemiol.* **37**(Suppl 1), 23–30, 2008.

Dunn W. B., Broadhurst D., Begley P., Zelena E., Francis-McIntyre S., Anderson N., Brown M., Knowles J. D., Halsall A., Haselden J. N., Nicholls A. W., Wilson

I. D., Kell D. B., and Goodacre R. Procedures for large-scale metabolic profiling of serum and plasma using gas ghromatography and liquid chromatography coupled to mass spectrometry. *Nat. Protoc.* **6**(7), 1060–1083, 2011a.

Dunn W. B., Broadhurst D. I., Atherton H. J., Goodacre R., and Griffin J. L. Systems level studies of mammalian metabolomes: the roles of mass spectrometry and nuclear magnetic resonance spectroscopy. *Chem. Soc. Rev.* **40**(1), 387–426, 2011b.

Dunn W. B., Erban A., Weber R. J. M., Creek D. J., Brown M., Breitling R., Hankemeier T., Goodacre R., Neumann S., Kopka J., and Viant M. Mass appeal: metabolite identification in mass mpectrometry-focused untargeted metabolomics. *Metabolomics*, 2012 In Press.

Eisenacher M. mzIdentML: an open community-built standard format for the results of proteomics spectrum identification algorithms. In *Data Mining in Proteomics* volume 696 of *Methods in Molecular Biology* pages 161–177. Humana Press, 2011.

Fahy E., Sud M., Cotter D., and Subramaniam S. LIPID MAPS online tools for lipid research. *Nucleic Acids Res.* **35**(suppl 2), W606–W612, 2007.

Faijes M., Mars A. E., and Smid E. J. Comparison of quenching and extraction methodologies for metabolome analysis of *Lactobacillus plantarum*. *Microb. Cell Fact.* **6**, 27, 2007.

Fairlamb A., Blackburn P., Ulrich P., Chait B., and Cerami A. Trypanothione: a novel bis(glutathionyl)spermidine cofactor for glutathione reductase in trypanosomatids. *Science* **227**(4693), 1485–1487, 1985.

Fan T. W. M., Lane A. N., Higashi R. M., Farag M. A., Gao H., Bousamra M., and Miller D. M. Altered regulation of metabolic pathways in human lung cancer discerned by ^{13}C stable isotope-resolved metabolomics (SIRM). *Mol. Cancer* **8**, 41, 2009.

Feist A. M., Herrgard M. J., Thiele I., Reed J. L., and Palsson B. Ø. Reconstruction of biochemical networks in microorganisms. *Nat. Rev. Microbiol.* **7**(2), 129–143, 2009.

Fiehn O. Metabolomics — the link between genotypes and phenotypes. *Plant Mol. Biol.* **48**(1), 155–171, 2002.

Fiehn O. Metabolite profiling in *Arabidopsis*. *Methods Mol. Biol.* **323**, 439–447, 2006.

Fink D., Falke D., Wohlleben W., and Engels A. Nitrogen metabolism in *Streptomyces coelicolor* A3(2): modification of glutamine synthetase I by an adenylyltransferase. *Microbiology* **145**(Pt 9), 2313–2322, 1999.

Garcia D. E., Baidoo E. E., Benke P. I., Pingitore F., Tang Y. J., Villa S., and Keasling J. D. Separation and mass spectrometry in microbial metabolomics. *Curr. Opin. Microbiol.* **11**(3), 233–239, 2008.

Giavalisco P., Köhl K., Hummel J., Seiwert B., and Willmitzer L. ^{13}C isotope-labeled metabolomes allowing for improved compound annotation and relative quantification in liquid chromatography-mass spectrometry-based metabolomic research. *Anal. Chem.* **81**(15), 6546–6551, 2009.

Gika H. G., Theodoridis G. A., Wingate J. E., and Wilson I. D. Within-day reproducibility of an HPLC–MS-based method for metabonomic analysis: application to human urine. *J. Proteome Res.* **6**(8), 3291–3303, 2007.

Gika H. G., Macpherson E., Theodoridis G. A., and Wilson I. D. Evaluation of the repeatability of ultra-performance liquid chromatography–TOF–MS for global metabolic profiling of human urine samples. *J. Chromatogr. B* **871**(2), 299–305, 2008.

Gipson G. T., Tatsuoka K. S., Sokhansanj B. A., Ball R. J., and Connor S. C. Assignment of MS-based metabolomic datasets via compound interaction pair mapping. *Metabolomics* **4**(1), 94–103, 2008.

Gonzalez-Salgado A., Steinmann M. E., Greganova E., Rauch M., Mäser P., Sigel E., and Bütikofer P. myo-Inositol uptake is essential for bulk inositol phospholipid but not glycosylphosphatidylinositol synthesis in *Trypanosoma brucei*. *J. Biol. Chem.* **287**(16), 13313–13323, 2012.

Granger J., Plumb R., Castro-Perez J., and Wilson I. Metabonomic studies comparing capillary and conventional HPLC-oa–TOF MS for the analysis of urine from Zucker obese rats. *Chromatographia* **61**, 375–380, 2005.

Halket J. M., Waterman D., Przyborowska A. M., Patel R. K., Fraser P. D., and Bramley P. M. Chemical derivatization and mass spectral libraries in metabolic profiling by GC/MS and LC/MS/MS. *J. Exp. Bot.* **56**(410), 219–243, 2005.

Hammond D. J. and Gutteridge W. E. UMP synthesis in the kinetoplastida. *Biochim. Biophys. Acta* **718**(1), 1–10, 1982.

Han J., Danell R. M., Patel J. R., Gumerov D. R., Scarlett C. O., Speir J. P., Parker C. E., Rusyn I., Zeisel S., and Borchers C. H. Towards high-throughput metabolomics using ultrahigh-field fourier transform ion cyclotron resonance mass spectrometry. *Metabolomics* **4**(2), 128–140, 2008.

Hebert C. G., Valdes J. J., and Bentley W. E. Beyond silencing–engineering applications of RNA interference and antisense technology for altering cellular phenotype. *Curr. Opin. Biotechnol.* **19**(5), 500–505, 2008.

Hegeman A. D., Schulte C. F., Cui Q., Lewis I. A., Huttlin E. L., Eghbalnia H., Harms A. C., Ulrich E. L., Markley J. L., and Sussman M. R. Stable isotope assisted assignment of elemental compositions for metabolomics. *Anal. Chem.* **79**(18), 6912–6921, 2007.

Heinemann M. and Sauer U. Systems biology of microbial metabolism. *Curr. Opin. Microbiol.* **13**(3), 337–343, 2010 <ce:title>Ecology and industrial microbiology ÃćâĆňÂć Special section: Systems biology</ce:title>.

Hiller K., Metallo C. M., Kelleher J. K., and Stephanopoulos G. Nontargeted elucidation of metabolic pathways using stable-isotope tracers and mass spectrometry. *Anal. Chem.* **82**(15), 6621–6628, 2010.

Hirumi H. and Hirumi K. Continuous cultivation of *Trypanosoma brucei* blood stream forms in a medium containing a low concentration of serum protein without feeder cell layers. *J. Parasitol.* **75**(6), 985–989, 1989.

Hopwood D. A. Streptomyces *in Nature and Medicine: The Antibiotic Makers*. Oxford University Press New York, 2007.

Horai H., Arita M., Kanaya S., Nihei Y., Ikeda T., Suwa K., Ojima Y., Tanaka K., Tanaka S., Aoshima K., Oda Y., Kakazu Y., Kusano M., Tohge T., Matsuda F., Sawada Y., Hirai M. Y., Nakanishi H., Ikeda K., Akimoto N., Maoka T., Takahashi H., Ara T., Sakurai N., Suzuki H., Shibata D., Neumann S., Iida T., Tanaka K., Funatsu K., Matsuura F., Soga T., Taguchi R., Saito K., and Nishioka T. MassBank: a public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* **45**(7), 703–714, 2010.

Horinouchi S. Mining and polishing of the treasure trove in the bacterial genus *Streptomyces*. *Biosci. Biotechnol. Biochem.* **71**(2), 283–299, 2007.

Imbert L., Ramos R. G., Libong D., Abreu S., Loiseau P. M., and Chaminade P. Identification of phospholipid species affected by miltefosine action in *Leishmania*

donovani cultures using LC–ELSD, LC–ESI/MS, and multivariate data analysis. *Anal. Bioanal. Chem.* **402**(3), 1169–1182, 2012.

Jaki B., Franzblau S., Cho S., and Pauli G. Development of an extraction method for mycobacterial metabolome analysis. *J. Pharm. Biomed. Anal.* **41**(1), 196–200, 2006.

Jankevics A., Merlo M. E., de Vries M., Vonk R. J., Takano E., and Breitling R. Metabolomic analysis of a synthetic metabolic switch in *Streptomyces coelicolor* A3(2). *Proteomics* **11**(24), 4622–4631, 2011.

Jankevics A., Merlo M. E., de Vries M., Vonk R. J., Takano E., and Breitling R. Separating the wheat from the chaff: a prioritisation pipeline for the analysis of metabolomics datasets. *Metabolomics* **8**(Suppl 1), 29–36, 2012.

Jebbar M., Sohn-Bösser L., Bremer E., Bernard T., and Blanco C. Ectoine-induced proteins in *Sinorhizobium meliloti* include an Ectoine ABC-type transporter involved in osmoprotection and ectoine catabolism. *J. Bacteriol.* **187**(4), 1293–1304, 2005.

Jones A. R. and Lister A. L. Managing experimental data using FuGE. *Methods Mol. Biol.* **604**, 333–343, 2010.

Kaliszan R. QSRR: quantitative structure-(chromatographic) retention relationships. *Chem. Rev.* **107**(7), 3212–3246, 2007.

Kaliszan R., Baczek T., Cimochowska A., Juszczyk P., Wiśniewska K., and Grzonka Z. Prediction of high-performance liquid chromatography retention of peptides with the use of quantitative structure-retention relationships. *Proteomics* **5**(2), 409–415, 2005.

Kamleh M. A., Barrett M. P., Wildridge D., Burchmore R. J. S., Scheltema R. A., and Watson D. G. Metabolomic profiling using Orbitrap Fourier transform mass spectrometry with hydrophilic interaction chromatography: a method with wide applicability to analysis of biomolecules. *Rapid Com. Mass Spectrom.* **22**(12), 1912–1918, 2008.

Kamleh M. A., Dow J. A. T., and Watson D. G. Applications of mass spectrometry in metabolomic studies of animal model and invertebrate systems. *Brief. Funct. Genomics Proteomics* **8**(1), 28–48, 2009a.

Kamleh M. A., Hobani Y., Dow J. A., Zheng L., and Watson D. G. Towards a platform for the metabonomic profiling of different strains of *Drosophila melanogaster* using liquid chromatography-Fourier transform mass spectrometry. *FEBS J.* **276**(22), 6798–6809, 2009b.

Kanehisa M., Goto S., Kawashima S., and Nakaya A. The KEGG databases at GenomeNet. *Nucleic Acids Res.* **30**(1), 42–46, 2002.

Kanehisa M., Goto S., Furumichi M., Tanabe M., and Hirakawa M. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.* **38**(Database issue), D355–360, 2010.

Kanehisa M., Goto S., Sato Y., Furumichi M., and Tanabe M. KEGG for integration and interpretation of large-scale molecular datasets. *Nucleic Acids Res.* **40**(Database issue), D109–114, 2012.

Karp P. D., Paley S., and Romero P. The Pathway Tools software. *Bioinformatics* **18**(Suppl. 1), S225–S232, 2002.

Karp P. D., Ouzounis C. A., Moore-Kochlacs C., Goldovsky L., Kaipa P., Ahrén D., Tsoka S., Darzentas N., Kunin V., and López-Bigas N. Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res.* **33**(19), 6083–6089, 2009.

Kasuya M. C., Cusi R., Ishihara O., Miyagawa A., Hashimoto K., Sato T., and Hatanaka K. Fluorous-tagged compound: a viable scaffold to prime oligosaccharide synthesis by cellular enzymes. *Biochem. Biophys. Res. Commun.* **316**(3), 599–604, 2004.

Katajamaa M. and Orešič M. Processing methods for differential analysis of LC/MS profile data. *BMC Bioinformatics* **6**, 179, 2005.

Kell D. B. Systems biology, metabolic modelling and metabolomics in drug discovery and development. *Drug Discov. Today* **11**(23-24), 1085–1092, 2006.

Keller B. O., Sui J., Young A. B., and Whittal R. M. Interferences and contaminants encountered in modern mass spectrometry. *Anal. Chim. Acta* **627**(1), 71–81, 2008.

Kempa S., Hummel J., Schwemmer T., Pietzke M., Strehmel N., Wienkoop S., Kopka J., and Weckwerth W. An automated GC×GC–TOF–MS protocol for batch-wise extraction and alignment of mass isotopomer matrixes from differential ^{13}C-labelling experiments: a case study for photoautotrophic-mixotrophic grown *Chlamydomonas reinhardtii* cells. *J. Basic Microbiol.* **49**(1), 82–91, 2009.

Kendall M. G. and Smith B. B. The problem of m rankings. *Ann. Math. Statistics* **10**, 275–287, 1939.

Keurentjes J. J., Fu J., de Vos C. H., Lommen A., Hall R. D., Bino R. J., van der Plas L. H., Jansen R. C., Vreugdenhil D., and Koornneef M. The genetics of plant metabolism. *Nat. Genet.* **38**(7), 842–849, 2006.

Khan M. M. and Martell A. E. Metal ion and metal chelate catalyzed oxidation of ascorbic acid by molecular oxygen. I. Cupric and ferric ion catalyzed oxidation. *J. Am. Chem. Soc.* **89**(16), 4176–4185, 1967.

Kind T. and Fiehn O. What are the obstacles for an integrated system for comprehensive interpretation of cross-platform metabolic profile data? *Bioanalysis* **1**(9), 1511–1514, 2009.

Kind T. and Fiehn O. Advances in structure elucidation of small molecules using mass spectrometry. *Bioanal. Rev.* **2**, 23–60, 2010.

Kol S., Merlo M. E., Scheltema R. A., de Vries M., Vonk R. J., Kikkert N. A., Dijkhuizen L., Breitling R., and Takano E. Metabolomic characterization of the salt stress response in *Streptomyces coelicolor*. *Appl. Environ. Microbiol.* **76**(8), 2574–2581, 2010.

Kueger S., Steinhauser D., Willmitzer L., and Giavalisco P. High-resolution plant metabolomics: from mass spectral features to metabolites and from whole-cell analysis to subcellular metabolite distributions. *Plant J.* **70**(1), 39–50, 2012.

Kuhl C., Tautenhahn R., Böttcher C., Larson T. R., and Neumann S. CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal. Chem.* **84**(1), 283–289, 2012.

Kuhn S., Helmus T., Lancashire R. J., Murray-Rust P., Rzepa H. S., Steinbeck C., and Willighagen E. L. Chemical markup, XML, and the World Wide Web. 7. CMLSpect, an XML vocabulary for spectral data. *J. Chem. Inf. Model.* **47**(6), 2015–2034, 2007.

Kuhn T., Willighagen E. L., Zielesny A., and Steinbeck C. CDK-Taverna: an open workflow environment for cheminformatics. *BMC Bioinformatics* **11**, 159, 2010 PMID: 20346188.

Lakshmanan V., Rhee K. Y., and Daily J. P. Metabolomics and malaria biology. *Mol. Biochem. Parasitol.* **175**(2), 104–111, 2011.

Lane A. N., Fan T. W., Bousamra M., Higashi R. M., Yan J., and Miller D. M. Stable isotope-resolved metabolomics (SIRM) in cancer research with clinical application to nonsmall cell lung cancer. *OMICS* **15**(3), 173–182, 2011.

Le A., Lane A. N., Hamaker M., Bose S., Gouw A., Barbi J., Tsukamoto T., Rojas C. J., Slusher B. S., Zhang H., Zimmerman L. J., Liebler D. C., Slebos R. J. C., Lorkiewicz P. K., Higashi R. M., Fan T. W. M., and Dang C. V. Glucose-independent glutamine metabolism via TCA cycling for proliferation and survival in b cells. *Cell Metab.* **15**(1), 110–121, 2012.

Lee D. Y. and Fiehn O. High quality metabolomic data for *Chlamydomonas reinhardtii*. *Plant Methods* **4**, 7, 2008.

Legendre P. Species associations: the Kendall coefficient of concordance revisited. *J. Agricult. Biol. Environ. Stat.* **10**, 226–245, 2005.

Letunic I., Yamada T., Kanehisa M., and Bork P. iPath: interactive exploration of biochemical pathways and networks. *Trends Biochem. Sci.* **33**(3), 101–103, 2008.

Lewis R. A., Shahi S. K., Laing E., Bucca G., Efthimiou G., Bushell M., and Smith C. P. Genome-wide transcriptomic analysis of the response to nitrogen limitation in *Streptomyces coelicolor* A3(2). *BMC Res. Notes* **4**, 78, 2011.

Lin H.-M., Edmunds S. J., Helsby N. A., Ferguson L. R., and Rowan D. D. Nontargeted urinary metabolite profiling of a mouse model of Crohn's disease. *J. Proteome Res.* **8**(4), 2045–2057, 2009.

Lioliou E., Romilly C., Romby P., and Fechter P. RNA-mediated regulation in bacteria: from natural to artificial systems. *New Biotechnol.* **27**(3), 222 – 235, 2010.

Lisec J., Schauer N., Kopka J., Willmitzer L., and Fernie A. R. Gas chromatography mass spectrometry-based metabolite profiling in plants. *Nat. Protocols* **1**, 387–396, 2006.

Lu W., Bennett B. D., and Rabinowitz J. D. Analytical strategies for LC–MS-based targeted metabolomics. *J. Chromatogr. B* **871**(2), 236–242, 2008.

Lu W., Clasquin M. F., Melamud E., Amador-Noguez D., Caudy A. A., and Rabinowitz J. D. Metabolomic analysis via reversed-phase ion-pairing liquid chromatography coupled to a stand alone Orbitrap mass spectrometer. *Anal. Chem.* **82**(8), 3212–3221, 2010.

Lumley T. and Miller A. *leaps: regression subset selection*. R package version 2.9, 2009.

Madsen R., Lundstedt T., and Trygg J. Chemometrics in metabolomics – a review in human disease diagnosis. *Anal. Chim. Acta* **659**(1–2), 23 – 33, 2010.

Maharjan R. P. and Ferenci T. Global metabolite analysis: the influence of extraction methodology on metabolome profiles of *Escherichia coli*. *Anal. Biochem.* **313**(1), 145–154, 2003.

Mal M., Koh P. K., Cheah P. Y., and Chan E. C. Y. Development and validation of a gas chromatography/mass spectrometry method for the metabolic profiling of human colon tissue. *Rapid Commun. Mass Spectrom.* **23**(4), 487–494, 2009.

Mallows C. L. Some comments on C_p. *Technometrics* **15**(4), 661—675, 1973.

Mannaert A., Downing T., Imamura H., and Dujardin J.-C. Adaptive mechanisms in pathogens: universal aneuploidy in *Leishmania*. *Trends Parasitol.* **28**(9), 370–376, 2012.

Matsuda F., Shinbo Y., Oikawa A., Hirai M. Y., Fiehn O., Kanaya S., and Saito K. Assessment of metabolome annotation quality: a method for evaluating the false discovery rate of elemental composition searches. *PLoS ONE* **4**(10), e7490, 2009.

Medema M. H., Breitling R., Bovenberg R., and Takano E. Exploiting plug-and-play synthetic biology for drug discovery and production in microorganisms. *Nat. Rev. Microbiol.* **9**(2), 131–137, 2011.

Melamud E., Vastag L., and Rabinowitz J. D. Metabolomic analysis and visualization engine for LC–MS data. *Anal. Chem.* **82**(23), 9818–9826, 2010.

Meyer H., Liebeke M., and Lalk M. A protocol for the investigation of the intracellular *Staphylococcus aureus* metabolome. *Anal. Biochem.* **401**(2), 250 – 259, 2010.

Mo M. L., Palsson B. Ø., and Herrgard M. J. Connecting extracellular metabolomic measurements to intracellular flux states in yeast. *BMC Syst. Biol.* **3**, 37, 2009.

Moco S., Vervoort J., Moco S., Bino R. J., Vos R. C. D., and Bino R. Metabolomics technologies and metabolite identification. *Trends Anal. Chem.* **26**(9), 855 – 866, 2007.

Mottram J. C., Robertson C. D., Coombs G. H., and Barry J. D. A developmentally regulated cysteine proteinase gene of *Leishmania mexicana*. *Mol. Microbiol.* **6**(14), 1925–1932, 1992.

Murakami T., Holt T. G., and Thompson C. J. Thiostrepton-induced gene expression in *Streptomyces lividans*. *J. Bacteriol.* **171**(3), 1459–1466, 1989.

Murphey B. F. and Nier A. O. Variations in the relative abundance of the carbon isotopes. *Phys. Rev.* **59**, 771–772, 1941.

Naderer T. and McConville M. J. The *Leishmania*–macrophage interaction: a metabolic perspective. *Cellular Microbiol.* **10**(2), 301–308, 2008.

Neumann S. and Böcker S. Computational mass spectrometry for metabolomics: identification of metabolites and small molecules. *Anal. Bioanal. Chem.* **398**, 2779–2788, 2010.

Nieselt K., Battke F., Herbig A., Bruheim P., Wentzel A., Jakobsen O., Sletta H., Alam M., Merlo M., Moore J., Omara W., Morrissey E., Juarez-Hermosillo M., Rodriguez-Garcia A., Nentwich M., Thomas L., Iqbal M., Legaie R., Gaze W., Challis G., Jansen R., Dijkhuizen L., Rand D., Wild D., Bonin M., Reuther J., Wohlleben W., Smith M., Burroughs N., Martin J., Hodgson D., Takano E., Breitling R., Ellingsen T., and Wellington E. The dynamic architecture of the metabolic switch in *Streptomyces coelicolor*. *BMC Genomics* **11**(1), 10, 2010.

Ogata H., Goto S., Sato K., Fujibuchi W., Bono H., and Kanehisa M. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **27**(1), 29–34, 1999.

Okubo Y., Yang S., Chistoserdova L., and Lidstrom M. E. Alternative route for glyoxylate consumption during growth on two-carbon compounds by Methylobacterium extorquens AM1. *J. Bacteriol.* **192**(7), 1813–1823, 2010.

Oldiges M., Lütz S., Pflug S., Schroer K., Stein N., and Wiendahl C. Metabolomics: current state and evolving methodologies and tools. *Appl. Microbiol. Biotechnol.* **76**(3), 495–511, 2007.

Olsen J. V., de Godoy L. M., Li G., Macek B., Mortensen P., Pesch R., Makarov A., Lange O., Horning S., and Mann M. Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. *Mol. Cell. Proteomics* **4**(12), 2010–2021, 2005.

Olszewski K. L., Mather M. W., Morrisey J. M., Garcia B. A., Vaidya A. B., Rabinowitz J. D., and Llinas M. Branched tricarboxylic acid metabolism in *Plasmodium falciparum*. *Nature* **466**(7307), 774–778, 2010.

Orth J. D., Thiele I., and Palsson B. Ø. What is flux balance analysis? *Nat. Biotechnol.* **28**(3), 245–248, 2010.

Park J. M., Kim T. Y., and Lee S. Y. Prediction of metabolic fluxes by incorporating genomic context and flux-converging pattern analyses. *Proc. Natl. Acad. Sci. U.S.A.* **107**(33), 14931–14936, 2010.

Pedrioli P. G. A., Eng J. K., Hubley R., Vogelzang M., Deutsch E. W., Raught B., Pratt B., Nilsson E., Angeletti R. H., Apweiler R., Cheung K., Costello C. E., Hermjakob H., Huang S., Julian R. K., Kapp E., McComb M. E., Oliver S. G., Omenn G., Paton N. W., Simpson R., Smith R., Taylor C. F., Zhu W., and Aebersold R. A common open representation of mass spectrometry data and its application to proteomics research. *Nat. Biotech.* **22**(11), 1459–1466, 2004.

Peyraud R., Kiefer P., Christen P., Massou S., Portais J. C., and Vorholt J. A. Demonstration of the ethylmalonyl-CoA pathway by using ^{13}C metabolomics. *Proc. Natl. Acad. Sci. U.S.A.* **106**(12), 4846–4851, 2009.

Pluskal T., Nakamura T., Villar-Briones A., and Yanagida M. Metabolic profiling of the fission yeast *S. pombe*: quantification of compounds under different temperatures and genetic perturbation. *Mol. Biosyst.* **6**(1), 182–198, 2010a.

Pluskal T., Castillo S., Villar-Briones A., and Orešič M. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **11**, 395, 2010b.

Price N. D., Reed J. L., and Palsson B. Ø. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat. Rev. Microbiol.* **2**(11), 886–897, 2004.

Prince J. T. and Marcotte E. M. Chromatographic alignment of ESI–LC–MS proteomics data sets by ordered bijective interpolated warping. *Anal. Chem.* **78**(17), 6140–6152, 2006.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. ISBN 3-900051-07-0, 2012.

Reaves M. L. and Rabinowitz J. D. Metabolomics in systems microbiology. *Curr. Opin. Biotechnol.* **22**(1), 17–25, 2011 <ce:title>Analytical biotechnology</ce:title>.

Reuther J. and Wohlleben W. Nitrogen metabolism in *Streptomyces coelicolor*: transcriptional and post-translational regulation. *J. Mol. Microbiol. Biotechnol.* **12**(1-2), 139–146, 2007.

Rijal S., Yardley V., Chappuis F., Decuypere S., Khanal B., Singh R., Boelaert M., Doncker S. D., Croft S., and Dujardin J.-C. Antimonial treatment of visceral leishmaniasis: are current *in vitro* susceptibility assays adequate for prognosis of *in vivo* therapy outcome? *Microbes Infect.* **9**(4), 529–535, 2007.

Rivière L., Moreau P., Allmann S., Hahn M., Biran M., Plazolles N., Franconi J.-M., Boshart M., and Bringaud F. Acetate produced in the mitochondrion is the essential precursor for lipid biosynthesis in procyclic trypanosomes. *Proc. Natl. Acad. Sci.* **106**(31), 12694–12699, 2009.

Rogers S., Scheltema R. A., Girolami M., and Breitling R. Probabilistic assignment of formulas to mass peaks in metabolomics experiments. *Bioinformatics* **25**(4), 512–518, 2009.

Rosenling T., Stoop M. P., Smolinska A., Muilwijk B., Coulier L., Shi S., Dane A., Christin C., Suits F., Horvatovich P. L., Wijmenga S. S., Buydens L. M., Vreeken R., Hankemeier T., van Gool A. J., Luider T. M., and Bischoff R. The impact of delayed storage on the measured proteome and metabolome of human cerebrospinal fluid. *Clin. Chem.* **57**(12), 1703–1711, 2011.

Rosenzweig D., Smith D., Opperdoes F., Stern S., Olafson R. W., and Zilberstein D. Retooling *Leishmania* metabolism: from sand fly gut to human macrophage. *FASEB J.* **22**(2), 590–602, 2008.

Rover L. J., Fernandes J. C., de Oliveira Neto G., Kubota L. T., Katekawa E., and Serrano S. H. Study of NADH stability using ultraviolet-visible spectrophotometric analysis and factorial design. *Anal. Biochem.* **260**(1), 50–55, 1998.

Ruiz B., Chávez A., Forero A., García-Huante Y., Romero A., Sánchez M., Rocha D., Sánchez B., Rodríguez-Sanoja R., Sánchez S., and Langley E. Production of microbial secondary metabolites: regulation by the carbon source. *Crit. Rev. Microbiol.* **36**(2), 146–167, 2010.

Salo M., Siren H., Volin P., Wiedmer S., and Vuorela H. Structure-retention relationships of steroid hormones in reversed-phase liquid chromatography and micellar electrokinetic capillary chromatography. *J. Chromatogr. A* **728**(1-2), 83–88, 1996.

Sangster T., Major H., Plumb R., Wilson A. J., and Wilson I. D. A pragmatic and readily implemented quality control strategy for HPLC–MS and GC–MS-based metabonomic analysis. *Analyst* **131**(10), 1075–1078, 2006.

Saunders E. C., De Souza D. P., Naderer T., Sernee M. F., Ralton J. E., Doyle M. A., Macrae J. I., Chambers J. L., Heng J., Nahid A., Likić V. A., and McConville M. J. Central carbon metabolism of *Leishmania* parasites. *Parasitology* **137**(9), 1303–1313, 2010.

Saunders E. C., Ng W. W., Chambers J. M., Ng M., Naderer T., Krömer J. O., Likić V. A., and McConville M. J. Isotopomer profiling of *Leishmania mexicana* promastigotes reveals important roles for succinate fermentation and aspartate uptake in tricarboxylic acid cycle (TCA) anaplerosis, glutamate synthesis, and growth. *J. Biol. Chem.* **286**(31), 27706–27717, 2011.

Saxena A., Lahav T., Holland N., Aggarwal G., Anupama A., Huang Y., Volpin H., Myler P., and Zilberstein D. Analysis of the *Leishmania donovani* transcriptome reveals an ordered progression of transient and permanent changes in gene expression during differentiation. *Molecular and Biochemical Parasitology* **152**(1), 53–65, 2007.

Scalbert A., Brennan L., Fiehn O., Hankemeier T., Kristal B. S., van Ommen B., Pujos-Guillot E., Verheij E., Wishart D., and Wopereis S. Mass-spectrometry-based metabolomics: limitations and recommendations for future progress with particular focus on nutrition research. *Metabolomics* **5**(4), 435–458, 2009.

Scheltema R. A., Kamleh A., Wildridge D., Ebikeme C., Watson D. G., Barrett M. P., Jansen R. C., and Breitling R. Increasing the mass accuracy of high–resolution LC–MS data using background ions — a case study on the LTQ-Orbitrap. *Proteomics* **8**(22), 4647–4656, 2008.

Scheltema R. A., Decuypere S., Dujardin J.-C., Watson D. G., Jansen R. C., and Breitling R. Simple data–reduction method for high-resolution LC–MS data in metabolomics. *Bioanalysis* **1**(9), 1551–1557, 2009.

Scheltema R. A., Decuypere S., t'kindt R., Dujardin J.-C., Coombs G. H., and Breitling R. The potential of metabolomics for *Leishmania* research in the post-genomics era. *Parasitology* **137**(9), 1291–1302, 2010.

Scheltema R. A., Jankevics A., Jansen R. C., Swertz M. A., and Breitling R. Peakml/mzMatch: a file format, Java library, R library, and tool-chain for mass spectrometry data analysis. *Anal. Chem.* **83**(7), 2786–2793, 2011.

Scherlach K. and Hertweck C. Triggering cryptic natural product biosynthesis in microorganisms. *Org. Biomol. Chem.* **7**(9), 1753–1760, 2009.

Sellick C. A., Hansen R., Maqsood A. R., Dunn W. B., Stephens G. M., Goodacre R., and Dickson A. J. Effective quenching processes for physiologically valid metabolite profiling of suspension cultured mammalian cells. *Anal. Chem.* **81**(1), 174–183, 2009.

Shah V. P., Midha K. K., Findlay J. W., Hill H. M., Hulse J. D., McGilveray I. J., McKay G., Miller K. J., Patnaik R. N., Powell M. L., Tonelli A., Viswanathan C. T., and Yacobi A. Bioanalytical method validation–a revisit with a decade of progress. *Pharm. Res.* **17**(12), 1551–1557, 2000 PMID: 11303967.

Silva A. M., Cordeiro-da Silva A., and Coombs G. H. Metabolic variation during development in culture of *Leishmania donovani* promastigotes. *PLoS Negl. Trop. Dis.* **5**(12), e1451, 2011.

Smith C. A., O'Maille G., Want E. J., Qin C., Trauger S. A., Brandon T. R., Custodio D. E., Abagyan R., and Siuzdak G. METLIN: a metabolite mass spectral database. *Ther. Drug Monit.* **27**(6), 747–751, 2005.

Smith C. A., Want E. J., O'Maille G., Abagyan R., and Siuzdak G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching and identification. *Anal. Chem.* **78**(3), 779–787, 2006.

Smith D. F., Peacock C. S., and Cruz A. K. Comparative genomics: from genotype to disease phenotype in the leishmaniases. *Int. J. Parasitol.* **37**(11), 1173 – 1186, 2007.

Smith T. K. and Bütikofer P. Lipid metabolism in *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* **172**(2), 66–79, 2010.

Snoep J. and Westerhoff H. From isolation to integration, a systems biology approach for building the silicon cell. In L. Alberghina and H. Westerhoff, editors, *Systems Biology* volume 13 of *Topics in Current Genetics* pages 13–30. Springer Berlin Heidelberg, 2005.

Sreekumar A., Poisson L. M., Rajendiran T. M., Khan A. P., Cao Q., Yu J., Laxman B., Mehra R., Lonigro R. J., Li Y., Nyati M. K., Ahsan A., Kalyana-Sundaram S., Han B., Cao X., Byun J., Omenn G. S., Ghosh D., Pennathur S., Alexander D. C., Berger A., Shuster J. R., Wei J. T., Varambally S., Beecher C., and Chinnaiyan

A. M. Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature* **457**(7231), 910–914, 2009.

Stoeckert C. J. and Parkinson H. The MGED ontology: a framework for describing functional genomics experiments. *Comp. Funct. Genomics* **4**(1), 127–132, 2003.

Sturm M. and Kohlbacher O. TOPPView: an open-source viewer for mass spectrometry data. *J. Proteome Res.* **8**(7), 3760–3763, 2009.

Sud M., Fahy E., Cotter D., Brown A., Dennis E. A., Glass C. K., Merrill A. H., Murphy R. C., Raetz C. R., Russell D. W., and Subramaniam S. LMSD: LIPID MAPS structure database. *Nucleic Acids Res.* **35**(Database issue), D527–532, 2007.

Sugimoto M., Hirayama A., Robert M., Abe S., Soga T., and Tomita M. Prediction of metabolite identity from accurate mass, migration time prediction and isotopic pattern information in CE–TOFMS data. *Electrophoresis* **31**(14), 2311–2318, 2010.

Sumner L. W., Amberg A., Barrett D., Beale M. H., Beger R., Daykin C. A., Fan T. W.-M., Fiehn O., Goodacre R., Griffin J. L., Hankemeier T., Hardy N., Harnly J., Higashi R., Kopka J., Lane A. N., Lindon J. C., Marriott P., Nicholls A., Reily M. D., Thaden J. J., and Viant M. R. Proposed minimum reporting standards for chemical analysis. *Metabolomics* **3**, 211–221, 2007.

Swertz M. A., van der Velde K. J., Tesson B. M., Scheltema R. A., Arends D., Vera G., Alberts R., Dijkstra M., Schofield P., Schughart K., Hancock J. M., Smedley D., Wolstencroft K., Goble C., de Brock E. O., Jones A. R., Parkinson H. E., and Jansen R. C. XGAP: a uniform and extensible data model and software platform for genotype and phenotype experiments. *Genome Biol.* **11**(3), R27, 2010.

Takano E., Chakraburtty R., Nihira T., Yamada Y., and Bibb M. J. A complex role for the gamma-butyrolactone SCB1 in regulating antibiotic production in *Streptomyces coelicolor A3(2). Mol. Microbiol.* **41**(5), 1015–1028, 2001.

Tang Y. J., Chakraborty R., Martin H. G., Chu J., Hazen T. C., and Keasling J. D. Flux analysis of central metabolic pathways in *Geobacter metallireducens* during reduction of soluble Fe(III)-nitrilotriacetic acid. *Appl. Environ. Microbiol.* **73**(12), 3859–3864, 2007.

Tang Y. J., Martin H. G., Myers S., Rodriguez S., Baidoo E. E., and Keasling J. D. Advances in analysis of microbial metabolic fluxes via ^{13}C isotopic labeling. *Mass Spectrom. Rev.* **28**(2), 362–375, 2009.

Tautenhahn R., Böttcher C., and Neumann S. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics* **9**, 504, 2008.

Taylor C. F., Paton N. W., Lilley K. S., Binz P.-A., Julian R. K. J., Jones A. R., Zhu W., Apweiler R., Aebersold R., Deutsch E. W., Dunn M. J., Heck A. J. R., Leitner A., Macht M., Mann M., Martens L., Neubert T. A., Patterson S. D., Ping P., Seymour S. L., Souda P., Tsugita A., Vandekerckhove J., Vondriska T. M., Whitelegge J. P., Wilkins M. R., Xenarios I., Yates, John R r., and Hermjakob H. The minimum information about a proteomics experiment (MIAPE). *Nat. Biotechnol.* **25**(8), 887–893, 2007 PMID: 17687369.

Thomason M. K. and Storz G. Bacterial antisense RNAs: how many are there, and what are they doing? *Annu. Rev. Genet.* **44**, 167–188, 2010.

Tiffert Y., Supra P., Wurm R., Wohlleben W., Wagner R., and Reuther J. The *Streptomyces coelicolor* GlnR regulon: identification of new GlnR targets and evidence for a central role of GlnR in nitrogen metabolism in actinomycetes. *Mol. Microbiol.* **67**(4), 861–880, 2008.

Tintaya K. W. Q., Ying X., Dedet J.-P., Rijal S., De Bolle X., and Dujardin J.-C. Antigen genes for molecular epidemiology of leishmaniasis: polymorphism of cysteine proteinase B and surface metalloprotease glycoprotein 63 in the *Leishmania donovani* complex. *J. Infect. Dis.* **189**(6), 1035–1043, 2004.

t'Kindt R., Scheltema R. A., Jankevics A., Brunker K., Rijal S., Dujardin J.-C., Breitling R., Watson D. G., Coombs G. H., and Decuypere S. Metabolomics to unveil and understand phenotypic diversity between pathogen populations. *PLoS Negl. Trop. Dis.* **4**(11), e904, 2010a.

t'Kindt R., Jankevics A., Scheltema R. A., Zheng L., Watson D. G., Dujardin J.-C., Breitling R., Coombs G. H., and Decuypere S. Towards an unbiased metabolic profiling of protozoan parasites: optimisation of a *Leishmania* sampling protocol for HILIC–Orbitrap analysis. *Anal. Bioanal. Chem.* **398**(5), 2059–2069, 2010b.

Trygg J., Holmes E., and Lundstedt T. Chemometrics in metabonomics. *J. Proteome Res.* **6**(2), 469–479, 2007.

van Weelden S. W., Fast B., Vogt A., van der Meer P., Saas J., van Hellemond J. J., Tielens A. G., and Boshart M. Procyclic *Trypanosoma brucei* do not use Krebs cycle activity for energy generation. *J. Biol. Chem.* **278**(15), 12854–12863, 2003.

Villas-Bôas S. G. and Bruheim P. Cold glycerol–saline: The promising quenching solution for accurate intracellular metabolite analysis of microbial cells. *Anal. Biochem.* **370**(1), 87 – 97, 2007.

Vincent I. M., Creek D. J., Burgess K., Woods D. J., Burchmore R. J., and Barrett M. P. Untargeted metabolomics reveals a lack of synergy between nifurtimox and eflornithine against *Trypanosoma brucei*. *PLoS Negl. Trop. Dis.* **6**(5), e1618, 2012.

Waters L. S. and Storz G. Regulatory RNAs in bacteria. *Cell* **136**(4), 615–628, 2009.

Watson D. G. The potential of mass spectrometry for the global profiling of parasite metabolomes. *Parasitology* **137**(9), 1409–1423, 2010.

Weber R. J. M. and Viant M. R. MI-Pack: increased confidence of metabolite identification in mass spectra by integrating accurate masses and metabolic pathways. *Chemom. Intell. Lab. Syst.* **104**(1), 75–82, 2010.

Weber T., Welzel K., Pelzer S., Vente A., and Wohlleben W. Exploiting the genetic potential of polyketide producing *Streptomycetes*. *J. Biotechnol.* **106**(2-3), 221–232, 2003.

Weininger D. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**(1), 31–36, 1988.

Whetzel P. L., Brinkman R. R., Causton H. C., Fan L., Field D., Fostel J., Fragoso G., Gray T., Heiskanen M., Hernandez-Boussard T., Morrison N., Parkinson H., Rocca-Serra P., Sansone S.-A., Schober D., Smith B., Stevens R., Stoeckert C. J., Taylor C., White J., and Wood A. Development of FuGO: an ontology for functional genomics investigations. *Omics* **10**(2), 199–204, 2006.

Winder C. L., Dunn W. B., Schuler S., Broadhurst D., Jarvis R., Stephens G. M., and Goodacre R. Global metabolic profiling of *Escherichia coli* cultures: an evaluation of methods for quenching and extraction of intracellular metabolites. *Anal. Chem.* **80**(8), 2939–2948, 2008.

Winder C. L., Dunn W. B., and Goodacre R. TARDIS-based microbial metabolomics: time and relative differences in systems. *Trends Microbiol.* **19**(7), 315–322, 2011.

Windig W. The use of the Durbin-Watson criterion for noise and background reduction of complex liquid chromatography/mass spectrometry data and a new algorithm to determine sample differences. *Chemom. Int. Lab. Syst.* **77**(1-2), 206–214, 2004.

Wishart D. S., Tzur D., Knox C., Eisner R., Guo A. C., Young N., Cheng D., Jewell K., Arndt D., Sawhney S., Fung C., Nikolai L., Lewis M., Coutouly M.-A., Forsythe I., Tang P., Shrivastava S., Jeroncic K., Stothard P., Amegbey G., Block D., Hau D. D., Wagner J., Miniaci J., Clements M., Gebremedhin M., Guo N., Zhang Y., Duggan G. E., MacInnis G. D., Weljie A. M., Dowlatabadi R., Bamforth F., Clive D., Greiner R., Li L., Marrie T., Sykes B. D., Vogel H. J., and Querengesser L. HMDB: the human metabolome database. *Nucleic Acids Res.* **35**(suppl 1), D521–D526, 2007.

Wishart D. S., Knox C., Guo A. C., Eisner R., Young N., Gautam B., Hau D. D., Psychogios N., Dong E., Bouatra S., Mandal R., Sinelnikov I., Xia J., Jia L., Cruz J. A., Lim E., Sobsey C. A., Shrivastava S., Huang P., Liu P., Fang L., Peng J., Fradette R., Cheng D., Tzur D., Clements M., Lewis A., De Souza A., Zuniga A., Dawe M., Xiong Y., Clive D., Greiner R., Nazyrova A., Shaykhutdinov R., Li L., Vogel H. J., and Forsythe I. HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res.* **37**(Database issue), D603–610, 2009.

Wittmann C., Krömer J. O., Kiefer P., Binz T., and Heinzle E. Impact of the cold shock phenomenon on quantification of intracellular metabolites in bacteria. *Anal. Biochem.* **327**(1), 135–139, 2004.

Wohlgemuth G., Haldiya P. K., Willighagen E., Kind T., and Fiehn O. The Chemical Translation Service–a web-based tool to improve standardization of metabolomic reports. *Bioinformatics* **26**(20), 2647–2648, 2010.

Wolf S., Schmidt S., Müller-Hannemann M., and Neumann S. *In silico* fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics* **11**, 148, 2010.

Woo H. M., Noack S., Seibold G. M., Willbold S., Eikmanns B. J., and Bott M. Link between phosphate starvation and glycogen metabolism in Corynebacterium glutamicum, revealed by metabolomics. *Appl. Environ. Microbiol.* **76**(20), 6910–6919, 2010.

Wu L., Mashego M. R., van Dam J. C., Proell A. M., Vinke J. L., Ras C., van Winden W. A., van Gulik W. M., and Heijnen J. J. Quantitative analysis of the microbial metabolome by isotope dilution mass spectrometry using uniformly ^{13}C-labeled cell extracts as internal standards. *Anal. Biochem.* **336**(2), 164–171, 2005.

Yamada T., Letunic I., Okuda S., Kanehisa M., and Bork P. iPath2.0: interactive pathway explorer. *Nucleic Acids Research* **39**(suppl 2), W412–W415, 2011.

Zamboni N. and Sauer U. Novel biological insights through metabolomics and ^{13}C-flux analysis. *Curr. Opin. Microbiol.* **12**(5), 553–558, 2009.

Zelena E., Dunn W. B., Broadhurst D., Francis-McIntyre S., Carroll K. M., Begley P., O'Hagan S., Knowles J. D., Halsall A., Wilson I. D., and Kell D. B. Development of a robust and repeatable UPLC–MS method for the long-term metabolomic study of human serum. *Anal. Chem.* **81**(4), 1357–1364, 2009.

Zheng L., t'Kind R., Decuypere S., von Freyend S. J., Coombs G. H., and Watson D. G. Profiling of lipids in *Leishmania donovani* using hydrophilic interaction chromatography in combination with Fourier transform mass spectrometry. *Rapid Commun. Mass Spectrom.* **24**(14), 2074–2082, 2010.

Samenvatting

Metabolomics stelt zich ten doel om biologie te bestuderen door algehele profilering van metabolieten: het gelijktijdig meten van zoveel mogelijk stoffen van laag molecu-lair gewicht in een biologisch monster. De favoriete techniek voor dit type profilering is vloeistofchromatografie gekoppeld aan massaspectrometrie (LC–MS).
Het doel van het werk beschreven in dit proefschrift was om open-source software te ontwikkelen voor de analyse van data gegenereerd door LC–MS-metabolomics. Een uiterst modulair ontwerp van de software stelde ons in staat om niet alleen de analytische componenten nauwkeurig af te stemmen op elk specifiek experimentele ontwerp, maar ook om op een intuïtieve manier de tussenliggende analyseresultaten te kunnen delen met de biologen en analytisch chemici die aan een project werken. Deze nieuwe aanpak maakt het mogelijk dat onderzoekers met verschillende achter-gronden goed met elkaar kunnen communiceren over de data, en heeft ertoe geleid dat bottlenecks in het ontwerp van experimenten, in de bioanalytische methodologie en in de data-analyse snel gelokaliseerd kunnen worden.
Het proefschrift begint met een inleiding over de voordelen van metabolomics met een massaspectrometrieplatform en een overzicht van de gehele experimentele pipeline (**Hoofdstuk 1**). In de **Hoofdstukken 2 t/m 6** worden verscheidene technieken voor datafiltering en metabolietidentificatie gepresenteerd. In **Hoofdstuk 7** geven we een voorbeeld van de flexibele computationele workflow voor data-analyse in actie. Deze methodologische hoofdstukken worden gevolgd door een reeks biologische stud-ies die metabolomics toepassen in de context van de systeembiologie. De optimalisatie van een algeheel metabolietenextractieprotocol voor *Leishmania donovani*-parasieten en de daaropvolgende optimalisatie van de analytische aanpak wordt beschreven in **Hoofdstuk 8**. In **Hoofdstuk 9** laten we zien dat de overexpressie van een antisense niet-coderend RNA gericht op glutamine synthetase I resulteert in een aanzienlijke reorganisatie van het metabolisme van de bacterie *Streptomyces coelicolor*, door meta-bolomics toe te passen met een nieuwe computationele aanpak die is gebaseerd op concordantie-analyse op een extreem groot aantal analytische replica's. **Hoofdstuk 10** laat de toepassing van stabiele isotopenlabeling met ongerichte metabolomics zien, teneinde een globaal overzicht te verkrijgen van het cellulaire lot van metabolieten

die benodigd zijn voor een bepaald metabolisch proces.
Het proefschrift wordt tenslotte afgesloten met **Hoofdstuk 11**, waarin we een toekomstbeeld schetsen van het metabolomics-onderzoek.

Kopsavilkums

Metabolomika kā zinātnes nozare ir viena no "omiku" tehnoloģijām, kas komplektā ar genomikas, transkriptomikas un proteomikas zināšanām nodrošina organismu funkciju norises izpratni, jo tā kompleksi pēta dzīva organisma vai šūnas metabolisma reakciju norisi laikā. Par metabolomu sauc šūnas, audu vai organisma visu metabolītu vai mazmolekulāro savienojumu kopumu noteiktos apstākļos. Metabolomu iegūst, ekstrahējot metabolītus no bioloģiskajiem šķidrumiem vai audiem, un to var uzskatīt par atbilstošo šūnu vai audu gēnu un proteīnu darbības rezultāta "pirkstu nospiedumu" vai signatūru.

Pēdējā laikā arī šķidrumu hromatogrāfija ar masselektīvo detektēšanu (LC–MS) ir attīstījusies par nozīmīgu instrumentu metaboloma pētījumos.

Šajā disertācijā ir apskatīta atvērtā koda programmatūras izstrāde LC–MS datu apstrādei. Izstrādātās programmas modulārā struktūra ļauj ne tikai dinamiski pielāgot dažādas komponentes katram pētījumam individuāli, bet arī nodrošina ērtu platformu starprezultātu apmaiņai starp projektā iesaistītajiem biologiem un analītiskajiem ķīmiķiem. Tādējādi tiek panākta daudz ērtāka un efektīvāka komunikācija starp atšķirīgu jomu zinātniekiem. Rezultātā ir iespējams daudz ātrāk noteikt un novērst nepilnības vai kļūdas eksperimenta struktūrā, analītiskajā platformā vai datu analīzē.

Disertācijas sākumā (**1. nodaļa**) ir apskatīti LC–MS metodes pielietojumi metabolomikas pētījumos kā arī datu apstrādes pamatprincipi. No **2. līdz 6. nodaļai** prezentētas vairākas perspektīvas stratēģijas datu filtrēšanai (priekšapstrādei) un metabolītu identificēšanai. **7. nodaļā** ir dots piemērs dinamiskās datu apstrādes darbplūsmas izmantošanai.

Nākošajās nodaļās prezentēti vairāki bioloģiskie pētījumi, kas demonstrē metabolomikas pielietojumu sistēmu bioloģijas kontekstā. Metabolītu ekstrahēšanas protokola optimizācija ar tai sekojošu analītiskās platformas optimizēšanu *Leishmania Donovano* parazītiem ir aprakstīta **8. nodaļā**. **9. nodaļā** demonstrējam, ka antiinoformācijas nekodējošās RNS, kas iedarbojas uz glutamīna sintetāzi I, pārekspresijas rezultātā ir novērojamas izteiktas izmaiņas baktērijas *Streptomyces Coelicolor* metabolomā. Šajā pētījumā tika izmantots ļoti liels skaits atkārtoto bioloģiko paraugu un jaunas datu apstrādes stratēģijas. **10. nodaļā**

demonstrējam smago izotopu iezīmētu savienojmu izmantošans iespējas metaboloma pētījumos, kas ļau izsekot prekursora savienojuma bioloģiskajām transformācijām. Disertācijas nobeigumā, **11. nodaļā** ir apskatītas metabolomikas izmantošanas iespējas un attīstības virzieni nākotnē.

Curriculum vitae

Andris Jankevics was born on May 11, 1981, in Riga, Latvia. In 2000 he began to study Chemistry at the University of Latvia and obtained a Bachelor degree in 2004. During his studies, he started to work as assistant (in organic synthesis) in the group of Dr. Pēteris Trapencieris in the Latvian Institute of Organic Synthesis. As Andris was increasingly interested in information technology-related subjects, after graduating, he went to work for the company Lintech as computer support manager, with emphasis on Linux-based systems and server environments. In 2006, he was offered an exciting opportunity to work in a completely new reasearch field in Latvia – Metabolomics. This interdisciplinary project (funded by the European Regional Development Fund) was supervised by Dr. Osvalds Pugovičs and As. Prof. Dr. Maija Dambrova and involved acquisition of metabolome data sets (on NMR and LC–MS platforms), data processing and statistical analysis. Andris continued his studies in the University of Latvia and obtained a Natural Sciences Master Degree in mathematics in 2008 and was awarded the Werner Siemens excellence prize for his thesis work. In 2009, he became a PhD candidate under the supervision of Prof. Dr. Rainer Breitling and Prof. Dr. Ritsert C. Jansen at the University of Groningen, working on the interpretation of challenging metabolomics data sets. He spent almost three years of his PhD as a visiting PhD student in the University of Glasgow, collaborating with several research groups across Europe and the UK. Since 2013, he is continuing to develop and improve LC–MS data processing algorithms at the University of Manchester, in the group of Prof. Dr. Rainer Breitling.

Selected conference presentations

June 25-28, 2012	**Metabolomics 2012 (Washington DC, USA)** Selected talk: *MzMatch/mzMatch.R: An open source software for the sequential processing and analysis of mass spectrometry data*
August 24-27, 2011	**Nordic Separation Science Society 6th Conference, (Riga, Latvia)** Invited talk: *Using LC-MS for Metabolic Systems Biology*
June 15-17, 2011	**EAST-NMR Regional Meeting "NMR and Complementary Tools for Chemistry and Biology Research" (Riga, Latvia)** Invited talk: *Metabolomic systems biology*
May 19-20, 2011	**Trends in Metabolomics – Analytics and Applications (Frankfurt, Germany)** Invited talk: *Separating the wheat from the chaff: a prioritisation pipeline for the analysis of metabolomics datasets*
February 15-19, 2010	**Seventh Winter Symposium on Chemometrics (Saint Petersburg, Russia)** Selected talk: *Quality control for large-scale LC–MS studies at runtime*

Seminars and workshops

- Co-organizer of the International Metabolomics Data Analysis Workshop
  - March 29-30, 2012, Glasgow, United Kingdom
  - November 21, 2011, Glasgow, United Kingdom
  - May 16, 2011, Glasgow, United Kingdom
  - November 30, 2010, Glasgow, United Kingdom
- Lectures in Chemometrics. Riga Technical University, Faculty of Material Sciences and Applied Chemistry, Riga, Latvia, Fall Semester, 2006-2011

Publications

Jankevics A, Gorkina J, Pugovics O, Vanaga A. Application of multivariate regression methods for HPLC determination of keratin and yeast content in pharmacological preparation. *Latvian Journal of Chemistry* 2007, 3:283–84

Jankevics A, Liepinsh E, Liepinsh E, Vilskersts R, Grinberga S, Pugovics O, Dambrova M. Metabolomic studies of experimental diabetic urine samples by ^{1}H NMR spectroscopy and LC/MS method. *Chemometrics and Intelligent Laboratory Systems* 2009, 97(1):11–17.

t'Kindt R, Scheltema RA, **Jankevics A**, Brunker K, Rijal S, Dujardin J-C, Breitling R, Watson DG, Coombs GH, Decuypere S. Metabolomics to unveil and understand phenotypic diversity between pathogen populations. *PLoS Neglected Tropical Diseases.* 2010, 4(11):e904.

t'Kindt R, **Jankevics A**, Scheltema RA, Zheng L, Watson DG, Dujardin J-C, Breitling R, Watson DG, Coombs GH, Decuypere S. Towards an unbiased metabolic profiling of protozoan parasites: Optimisation of a *Leishmania* sampling protocol for HILIC–Orbitrap analysis. *Analytical and Bioanalytical Chemistry* 2010, 398(5):2059–69.

Scheltema RA*, **Jankevics A***, Jansen RC, Swertz MA, Breitling R. PeakML/mzMatch: A file format, java library, R library, and tool-chain for mass spectrometry data analysis. *Analytical Chemistry* 2011, 83(7):2786–93.

Merlo ME, **Jankevics A**, Takano E, Breitling R. Exploring the metabolic state of microorganisms using metabolomics. *Bioanalysis* 2011, 3(21):2443–58.

Creek DJ, **Jankevics A**, Breitling R, Watson DG, Barrett MP, Burgess KEV. Toward global metabolomics analysis with hydrophilic interaction liquid chromatography–mass spectrometry: Improved metabolite identification by retention time prediction. *Analytical Chemistry* 2011, 83(22):8703–10.

Jankevics A*, Merlo ME*, de Vries M, Vonk RJ, Takano E, Breitling R. Metabolomic analysis of a synthetic metabolic switch in *Streptomyces coelicolor* A3(2). *Proteomics* 2011, 11(24):4622–31.

Creek DJ, **Jankevics A**, Burgess KEV, Breitling R, Barrett MP. IDEOM: An excel interface for analysis of LC–MS-based metabolomics data. *Bioinformatics* 2012, 28(7):1048–9.

Jankevics A, Merlo ME, de Vries M, Vonk RJ, Takano E, Breitling R. Separating the wheat from the chaff: A prioritisation pipeline for the analysis of metabolomics datasets. *Metabolomics* 2012, 8(Supp. 1):29–36.

Nguyen Q-T, Merlo ME, Medema MH, **Jankevics A**, Breitling R, Takano E. Metabolomics methods for the synthetic biology of secondary metabolism. *FEBS Letters* 2012, 586(15):2177–83.

Creek DJ, Chokkathukalam A, **Jankevics A**, Burgess KEV, Breitling R, Barrett MP. Stable isotope-assisted metabolomics for network-wide metabolic pathway elucidation. *Analytical Chemistry* 2012, 84(20):8442–8447.

Chaneton B, Hillmann P, Zheng L, Martin ACL, Maddocks ODK, Chokkathukalam, A, Coyle JE, **Jankevics A**, Holding FP, Vousden KH, Frezza C, O'Reilly M, Gottlieb E. Serine is a natural ligand and allosteric activator of pyruvate kinase M2. *Nature* 2012, 491(7424):458–62.

Chokkathukalam A*, **Jankevics A***, Creek DJ, Achcar F, Barrett MP, Breitling R. mzMatch–ISO: an R tool for the annotation and relative quantification of isotope-labelled mass spectrometry data. *Bioinformatics* 2013, 29(2):281–283.

Berg M, Vanaerschot M, **Jankevics A**, Cuypers B, Breitling R, Dujardin J-C. LC–MS metabolomics from study design to data-analysis — using a versatile pathogen as a test case. *Computational and Structural Biotechnology Journal* 2013, 4(5):e201301002.

* – equal contribution

Acknowledgements

First of all, I would like to acknowledge my supervisors, Prof. Rainer Breitling and Prof. Ritsert Jansen. I learned a lot from both of you and I strongly appreciate the comments, constructive criticism and encouragement that you provided me. Rainer, thanks a lot for giving me a chance to work with you, and for your support while preparing this manuscript.

I wish to express my deepest gratitude to our collaborators, without whom this thesis would not have been possible. They are Prof. Eriko Takano, Dr. Elena Merlo, ing. Marcel de Vries, Quoc-Thai Nguyen, Prof. Roel J. Vonk, Prof. Michael P. Barrett, Dr. Darren J. Creek, Dr. Karl E. V. Burgess, Dr. Gavin Blackburn, Prof. Jean-Claude Dujardin, Dr. Saskia Decuypere, Dr. Ruben t'Kindt, Dr. Maya Berg and Dr. Manu Vanaerschot. I am deeply grateful to Saskia, Ruben and Maya, who never grew tired of testing new algorithms and software and who supplied me with lot of great ideas and suggestions. Dear Elena, thank you for your input in generating loads of challenging data sets, good work discussions and friendly chats during coffee breaks. Darren, Karl and Gavin, I really appreciate your input and the fruitful discussions we had.

Dear Richard, it was fun to work with you. I enjoyed our conversations during coffee breaks and the great inspiration you gave me when I started my project in Groningen.

No less important, I am deeply grateful to the secretaries of GbiC Klazien and to Ingrid for the help in solving issues encountered during my PhD.

Dear Elena and Richard, thanks again for agreeing to be my paranymphs.

I am grateful to all my colleagues for the daily company, technical support, good advice and interesting chats. Current and former members of GbiC in Groningen: Tauqeer, Bruno, Marnix, Yang, Danny, Joeri, Rene, Nino, Morris, Frank, George and Anna. After my move to Glasgow, I enjoyed working together with Unni, Fiona, Rónán, Jagtar, Eduard and Roberta. A special thanks to Marnix for translating the thesis summary in to Ducth. Rónán, thank you for your assistance with the programming in Java. Dear Unni, thank you for joining me during those long cycle

rides across Scotland, for the encouragement, for always being a good friend and colleague and for the support you provided me during the final years of my PhD project.
I thank all my friends who I met in Groningen and later also in Glasgow. A special thanks to Sophia for your permission to use one of your photos on my thesis book cover, and for being of great company for going outdoors, attending gigs, local coffee shops, etc.
I am thankful to Prof. Johan Westerhuis and Prof. Matthias Heinemann for sending me valuable comments on my thesis during the thesis approval process.
Visbeidzot es vēlētos izteikt vissirsnīgāko pateicību mammai, Annai un Dacei. Es augstu novērtēju jūsu sniegto palīdzību un atbalstu, bez kura es nebūtu sasniedzis savus mērķus. Šo darbu vēlos veltīt sava Tēva gaišajai piemiņai.

Andris
Manchester
July 25, 2013